Leaving the Cathedral

Ryan Faulk

August 6, 2020

Contents

1 Introduction **3**

2 Predictions **4**
2.1 "Superforecasters" . . . 4
2.2 Statistical Prediction Rules . . . 5
2.3 Financial Predictions . . . 7
2.4 Behavioral Economists vs. Mechanical Turk . . . 8
2.5 Law Cases . . . 10

3 Not Supermen **13**
3.1 IQ by Education Level . . . 13
3.2 Knowledge Decay . . . 16
3.3 Bullying . . . 20
3.4 On Psychology . . . 26
3.5 Knowledge of Statistics . . . 28
3.5.1 McShane 2016 . . . 29
3.5.2 McShane 2017 . . . 32
3.5.3 Lyu 2019 . . . 33
3.5.4 Zuckerman 1993 . . . 36
3.5.5 Hoekstra 2014 . . . 36
3.5.6 Haller 2002 . . . 37
3.6 Nonsense Math . . . 38
3.6.1 Study 1 . . . 39
3.6.2 Study 2 . . . 40
3.6.3 Study 3 . . . 41

4 The Journal System **41**
4.1 Word Game . . . 41
4.2 Basic Knowledge Problem . . . 42
4.3 Big vs. Small Journals . . . 42
4.4 Article Prominence . . . 47
4.5 Outcome Stings . . . 48
4.6 BMJ Error Detection Sting . . . 51
4.7 Prestige Stings . . . 52

4.8 Fake Papers 55

5 The Great Stagnation **57**
5.1 Economic "stagnation" 57
5.2 Technological Stagnation 62
5.3 Stagnation occurred in multiple fields at the same time 67
5.4 This stagnation occurred with the rise of the journal system and the "corporatization" or "institutionalization" of academia. 69
5.5 There are intuitive causal ways that the changes in "institutional science" can cause a decline 70
5.6 Explosion in the number of PhDs 71
5.7 A Decline in "g" 72
5.8 Demographic changes 76
5.9 Incentive Problem 80
5.10 Origin of the University 81
5.11 Organizations 82
5.12 Scam 83

1 Introduction

This first section on authorities is difficult because it makes us look like cranks. Who tends to question the institutions which define truth at a given place and time in history? Well, people who, for whatever reason, tend to distrust the authorities on a particular issue. Today, this can be people who believe in UFO abductions, urine therapy, bigfoot, remote viewing, telekinesis, ancient aliens, and lots of other things that appear to be nonsense to the typical person.

And it's a cycle. What you believe is true is defined by your historically contextual authorities, as a kind of default. Maybe not 100%, but you default to the authoritative position on any given issue. That's not necessarily a bad thing - the world is a big place and you can't know everything about everything and you have to outsource to experts, or who you believe to be experts, sometimes. We don't believe in UFO abductions, but at the same time, we don't see it as any more absurd than believing evolution stopped at the neck.

The same kind of effect is seen in Scientology v. Islam. Sure, we think Scientology and Xenu is absurd, but it's no more absurd than Mohammed and an angel flying him around the world. Muslims would disagree.

But the problem then is that anyone who is making a fundamental critique of authority (and lets be real, today that means academic authority and really nothing else) almost always holds some "absurd" view. A biblical creationist can be correct when he says scientists generally don't give biblical creationism a fair shake. He can be correct in finding flaws in some evolutionary narratives, perhaps because he's the only one looking for them. Which is to say, his critique of power can be absolutely correct even if it's motivated by a desire to advance something which is wrong - assuming biblical creationism is wrong.

And that's the cycle. Critiques of power tend to come from people who don't have power, and thus are effectively pathologized as kooks and cranks. But even if these "kooky" and "cranky" ideas are wrong, that doesn't mean the critique of power is wrong. And this is the problem; the fact that critiques of power tend to come from "kooks and cranks" results in any critique of power being pathologized. So, lets just go ahead and stop doing that alright?

Isaac Newton believed in astrology and his principal focus was on alchemy. Lets not use that to hand-wave away the *Principia* because it came from an astrology-believing alchemist "crank."

The old church, the press, corporations, don't really have any authority. Media reports have the problem of being trusted by default, however they're not a barrier to intellectual change. They are not immune to challenge.

When we have talked to people about race differences in intelligence for example, nobody brings up some Vox article as a rebuttal on the basis of authority. They may bring it up as an argument, which stands or falls on it's own, but nobody is citing Vox *as an authority* - that you should be expected to believe this simply because some writer from Vox said so.

The only institution that is brought up on purely authoritarian grounds is the University Credential system. You should believe X because, supposedly, these people believe X. Now even that is often dubious because it's usually not based on surveys, and if it is, well, all the survey technically says is what someone is saying on a survey.

And we're not idiots or chumps. While "appeal to authority" is a "fallacy," and supposedly we could just say that and be done with this - nobody actually cares about fallacies. And while we don't care about academic consensus on anything *inherently*, we know you care. We care because you care. And so the first section is dedicated to academic authority.

But don't infer anything not explicitly said here. We're not conceding anything. Not even

academic authority, which we have more of than you might imagine. But that will be discussed at it's own length. For now, lets dive on in.

2 Predictions

2.1 "Superforecasters"

In his book *Expert Political Judgment,* Philip Tetlock looked at a sample of 177 "superforecasters." These are people who predict all manner of political events that are put on political betting markets, and are able to make money doing so.

Examples of things they bet on would be elections, but also things like the extent of the Arctic and Antarctic ice sheet, how many troops the US will have in Iraq or Afghanistan by a certain date, et cetera.

And these people make objective predictions which are scored. They can't be vague predictions like "the US will pull out a lot of troops," one has to specify how many, and by what date. And scores have to be standardized across different topics. For example comparing number of troops to square kilometers of Arctic ice - the degree of variance in those things have to be standardized so that a predictor can be scored on the same scale for both topics. How well someone does with their predictions is called their "Brier Score." Forecasters with a Brier score above a certain value labeled by Tetlock "superforecasters."

TABLE 3.1
Individual Difference Predictors of Calibration of Subjective Probability Forecasts

	Forecasting Accuracy	
Individual Difference Predictors	*Correlations*	*Standardized Betas (with standard errors)*
I. Professional Background		
(a) Education (Ph.D. or not)	+.02	+.001 (.03)
(b) Years of professional experience (1 to 36)	+.00	+.02 (.03)
(c) Academic or nonacademic work	−.03	+.05 (.04)
(d) Access to classified information	+.02	+.01 (.05)
(e) Contact with media (1–7 scale: never to every week)	−.12	−.09 (.08)
Gender (female = 1)	.05	.08 (.08)
(f) Self-rate relevance of expertise	.09	.03 (.07)
II. Ideological-Theoretical Orientation		
(a) Left-Right	+.07	+.01 (.05)
(b) Idealist-Realist	+.06	−.03 (.06)
(c) Doomster-Boomster	+.20*	−.12 (.04)*
III. Cognitive Style		
(a) Hedgehog-Fox	.35**	+.29 (.04)**
(b) Integratively complex thought protocols	.31	+.25 (.05)**
(c) Extremism	.30	+.09 (.06)

* .05 significance
** .01 significance
Adjusted R^2 = .29 (N = 177)

What Tetlock found was illuminating. Education had no effect, years of professional experience had no effect, being an academic or non-academic had no effect. Contact with media had a small

(-.12) effect, with greater media exposure resulting in less accurate predictions. Females were 5% more accurate, and the correlation between self-rated expertise and predictive accuracy was .09.

"Left-Right" political orientation had virtually no effect, nor did seeing yourself as an "idealist" or "realist." The doomster-boomster dichotomy did have a significant effect, with boomsters (economic and environmental optimists) predicting with 20% greater accuracy.

The point of the book was on the "hedgehog vs. fox" thinking style, which is that the fox knows about a lot of little things, while the hedgehog knows one big thing. Which is to say, the "hedgehog" thinker has a more coherent overarching worldview; a "hedgehog" on economic questions would be a Marxist or a Friedmanite free marketer. And "foxes" were 35% more accurate than hedgehogs, a huge effect. Those with more extreme predictions made better predictions, and those with more "integratively complex thought protocols" made better predictions - which means people who had a system of formal rules in their thinking did better than those who didn't.

And Tetlock's book is very interesting, and very little of his book is about highlighting the irrelevance of formal expertise. He mentions it in passing. But while not Tetlock's focus, that is something incredibly important to society - that among the "superforecasters," credential is virtually irrelevant. Even when the credential is on the topic being predicted, and the forecaster rates his expertise highly as a result, the increase in accuracy of prediction compared to the rest of the forecasters is only 9%. Which is a smaller effect than just being an optimist or not having your brain rotted by news media.

So be more of an economic and ecological optimist, stop watching the news media, be less ideological (less of a hedgehog and more of a fox), develop some formal rules of thinking (or at the very least sound out your arguments), and that will do far more for your ability to predict the world in a general sense than a credential will, which at most will mildly increase the accuracy of your predictions in a narrow set of topics. You may want a credential for other reasons, but it's not going to make you more knowledgeable about the world in any testable way on difficult and contested questions.

Now this is not to say that Non-PhDs are just as good at predicting things as PhDs. They may or may not be. We would guess PhDs would be better at predicting things. The data on superforecasters merely compares non-PhD *superforecasters* to PhD *superforecasters*; that is, among the population of superforecasters, having a PhD doesn't matter.

PhDs may (literally - may) be more likely to reach the threshold or superforecasters, but once you pass that bar, it doesn't matter.

2.2 Statistical Prediction Rules

An old paper, from the year 2000, looked at comparing statistical prediction rules vs. field-relevant experts, and compared the standardized accuracy of these prediction rules vs. certified clinical experts.

This is from the paper "Clinical Versus Mechanical Prediction: A Meta-Analysis"[1], the researchers looked at 136 studies with that comparison. The categories were Educational, Financial, Forensic, Medical, and Clinical - Personality. And in all categories the "mechanical prediction" outperformed clinical prediction.

The authors then organized the results into 5 categories:

[1] https://sci-hub.tw/10.1037/1040-3590.12.1.19

Table 1
Studies Included in Meta-Analysis

Citation	Predictand	Accuracy statistic	Accuracy	
			Clinical	Mechanical
Alexakos (1966)	college academic performance	HR	39	56
Armitage & Pearl (1957)	psychiatric diagnosis	HR	30	31
Ashton (1984)	*magazine advertising sales*	*corr*	0.63	0.88
Barron (1953)	psychotherapy outcome	HR	62	73
Blattberg & Hoch (1988)	catalog sales; coupon redemption	corr	0.52	0.66
Blenkner (1954)	case work outcome	corr	0.00	0.62
Bobbitt & Newman (1944)	success in military training	regression coefficient	0.93	0.87
Bolton et al. (1968)	vocational rehabilitation outcome	corr	0.30	0.40
Boom (1986)	diagnosis of jaundice	HR	85	90
Boom et al. (1988)	diagnosis of jaundice	HR	88	96
Boyle et al. (1966)	diagnosis of thyroid disorder	HR	77	85
Brodman et al. (1959)	general medical diagnosis	HR	43	48
Brown et al. (1989)	diagnosis of lateralized cerebral dysfunction	corr	0.43	0.64
Buss et al. (1955)	prediction of anxiety	corr	0.60	0.64
Caceres & Hochberg (1970)	diagnosis of heart disease	HR	74	84
Campbell et al. (1962)	job performance	corr	0.15	0.29
Cannon & Gardner (1980)	general medical diagnoses, optimality of treatment recommendations	HR	63	64
Cebul & Poses (1986)	presence of throat infection	HR	69	99
Clarke (1985)	surgery recommendation	HR	59	69
Cooke (1967)	psychological disturbance	HR	77	76
Cornelius & Lyness (1980)	job analysis	corr	0.73	0.76
Danet (1965)	future psychiatric illness	HR	65	70
Dannenberg et al. (1979)	prognosis of medical illness	accuracy coefficient	0.22	0.21
Dawes (1971)	success in graduate school	corr	0.10	0.51
De Dombal et al. (1974)	diagnosis of gastrointestinal disorders	HR	71	92
De Dombal et al. (1975)	diagnosis of gastrointestinal disorders	HR	83	85
De Dombal, Horrocks, et al. (1972)	diagnosis of gastrointestinal disorders	HR	50	97
De Dombal, Leaper, et al. (1972)	diagnosis of appendicitis	*HR*	83	92
Devries & Shneidman (1967)	course of psychiatric symptoms	HR	75	100
Dicken & Black (1965)	supervisory potential	corr	0.09	0.30
Dickerson (1958)	client compliance with counseling plan	HR	57	52
Dickson et al. (1985)	diagnosis of abdominal pain	HR	55	73
Dunham & Meltzer (1946)	length of psychiatric hospitalization	HR	34	70
Dunnette et al. (1960)	job turnover	HR	53	73
Durbridge (1984)	diagnosis of hepatic or biliary disorder	HR	62	74
Edwards & Berry (1974)	psychiatric diagnosis	HR	63	74
Enenkel & Spiel (1976)	diagnosis of myocardial infarction	HR	78	57
Evenson et al. (1973)	medication prescribed	HR	77	75
Evenson et al. (1975)	length of hospitalization	HR	76	71
Geddes et al. (1978)	degree of pulmonary obstruction	HR	96	95
Glaser & Hangren (1958)	probation success	*HR*	83	84
Glaser (1955)	criminal recidivism	mean cost rating	0.14	0.35
S. C. Goldberg & Mattsson (1967)	improvement of schizophrenia	significance test	8.15	10.78
L. R. Goldberg (1965)	psychiatric diagnosis	corr	0.28	0.38
L. R. Goldberg (1969)	psychiatric diagnosis	HR	62	69
L. R. Goldberg (1976)	business failure	corr	0.51	0.56
Goldman et al. (1981)	cardiac disease survival or remission	corr	−0.12	−0.11
Goldman et al. (1982)	diagnosis of acute chest pain	HR	79	73
Goldman et al. (1988)	prediction of myocardial infarction	HR	73	76
Goldstein et al. (1973)	cerebral impairment	HR	95	75
Gottesman (1963)	personality description	HR	62	53
Grebstein (1963)	prediction of IQ	corr	0.59	0.56
Gustafson et al. (1973)	diagnosis of thyroid disorder	HR	88	87
Gustafson et al. (1977)	suicide attempt	HR	63	81
Halbower (1955)	personality description	corr	0.42	0.64
Hall (1988)	criminal behavior	HR	54	83
Hall et al. (1971)	diagnosis of rheumatic heart disease	HR	62	73
Harris (1963)	game outcomes and point spread	HR	60	69
Hess & Brown (1977)	academic performance	HR	68	83
Holland et al. (1983)	criminal recidivism	corr	0.32	0.34
Hopkins et al. (1980)	surgical outcomes	HR	84	91
Hovey & Stauffacher (1953)	personality characteristics	HR	74	63
Ikonen et al. (1983)	diagnosis of abdominal pain	HR	67	59
Janzen & Coe (1973)	"diagnosis" of female homosexuality	HR	57	85
Jeans & Morris (1976)	diagnosis of small bowel disease	HR	83	83
Johnston & McNeal (1967)	length of psychiatric hospitalization	HR	72	75
Joswig et al. (1985)	diagnosis of recurrent chest pain	HR	69	86
Kahn et al. (1988)	detection of malingering	HR	21	25
Kaplan (1962)	psychotherapy outcome	HR	66	70
Kelly & Fiske (1950)	success on psychology internship	corr	0.32	0.41
Khan (1986)	business startup success	corr	−0.09	0.13
Klehr (1949)	psychiatric diagnosis	HR	67	64
Klein et al. (1973)	psychopharmacologic treatment outcome	corr	0.12	0.90
Kleinmuntz (1963)	maladjustment	HR	70	72
Kleinmuntz (1967)	maladjustment	HR	68	75
Klinger & Roth (1965)	diagnosis of schizophrenia	HR	77	43
Kunce & Cope (1971)	job success	HR	67	77
Lee et al. (1986)	death and myocardial infarction	corr	0.58	0.64
Leli & Filskov (1981)	presence, chronicity and lateralization of cerebral impairment	HR	79	79
Leli & Filskov (1984)	diagnosis of intellectual deterioration	HR	75	73
Lemerond (1977)	suicide	HR	50	50
Lewis & MacKinney (1961)	career satisfaction	corr	0.09	0.56
Libby (1976)	business failure	HR	74	72
Lindzey (1965)	"diagnosis" of homosexuality	HR	70	57
Lindzey et al. (1958)	"diagnosis" of homosexuality	HR	95	85
Lyle & Quast (1976)	diagnosis of Huntington disease	HR	61	68
Martin et al. (1960)	diagnosis of jaundice	HR	87	79
Mathew et al. (1988)	diagnosis of low back pain	HR	74	87
McClish & Powell (1989)	intensive care unit mortality	ROC	0.89	0.83
Miller et al. (1982)	general medical diagnosis	HR	53	40
Mitchell (1975)	managerial success	corr	0.19	0.46
Oddie et al. (1974)	diagnosis of thyroid disorder	HR	97	99
Orient et al. (1985)	diagnosis of abdominal pain	HR	64	63
Oskamp (1962)	presence of psychiatric symptoms	HR	70	71
Peck & Parsons (1956)	work productivity	corr	0.71	0.61
Pierson (1958)	college success	HR	43	49
Pipberger et al. (1975)	diagnosis of cardiac disease	HR	72	91
Plag & Weybrew (1968)	fitness for military service	corr	0.19	0.30
Popovics (1983)	cerebral dysfunction	corr	0.17	0.16
Poretsky et al. (1985)	diagnosis of myocardial infarction	HR	80	67
Reale et al. (1968)	diagnosis of congenital heart disease	HR	73	82
Reich et al. (1977)	diagnosis of hematologic disorders	HR	68	71
Reitan et al. (1964)	diagnosis of cerebral lesions	HR	75	73
Rosen & Van Horn (1961)	academic performance	HR	55	57
Royce & Weiss (1975)	marital satisfaction	corr	0.40	0.58
Sacks (1977)	criminal recidivism	HR	72	78
Sarbin (1942)	academic performance	corr	0.35	0.45
Schiedt (1936)	parole success or failure	HR	68	76
Schofield & Garrard (1975)	performance in medical school	HR	76	78
Schofield (1970)	performance in medical school	deviation score	0.07	−0.06
Schreck et al. (1986)	diagnosis of acid-base disorders	HR	55	100
Schwartz et al. (1976)	diagnosis of metabolic illnesses	HR	92	85
Shapiro (1977)	outcome of rheumatic illness	Q	0.20	0.15
Silverman & Silverman (1962)	diagnosis of schizophrenia	HR	55	64
Smith & Lanyon (1968)	juvenile criminal recidivism	HR	52	54
Speigelhalter & Knill-Jones (1984)	diagnosis of dyspepsia	ROC	0.85	0.83
Stephens (1970)	schizophrenia prognosis and course	corr	0.51	0.29
Storment & Finney (1953)	assaultive behavior	corr	0.00	0.57
Sutton (1989)	diagnosis of abdominal pain	HR	65	57
Szucko & Kleinmuntz (1981)	lie detection	corr	0.23	0.42
Taulbee & Sisson (1957)	psychiatric diagnosis	HR	63	63
Thompson (1952)	juvenile delinquency	HR	64	91
Truesdell & Bath (1957)	academic dropouts	HR	71	75
Ullman (1958)	course of group home placement	HR	59	78
Walters et al. (1988)	malingering	HR	56	93
Warner (1964)	diagnosis of congenital heart disease	HR	66	66
Watley & Vance (1963)	college achievement and leadership	HR	59	72
Webb et al. (1975)	occupational choice	HR	35	55
Wedding (1983)	diagnosis of cerebral impairment	corr	0.74	0.84
Weinberg (1957)	personality characteristics	corr	0.41	0.65
Werner et al. (1984)	assault by psychiatric inpatients	corr	0.14	0.56
Wexler et al. (1975)	medical diagnosis	HR	65	85
Wiggins & Kohen (1971)	graduate school success	corr	0.33	0.58
Wilkinson & Markus (1989)	minor psychiatric morbidity	ROC	0.74	0.89
Wittman & Steinberg (1944)	psychiatric prognosis	HR	41	68
Wormith & Goldstone (1984)	criminal recidivism	corr	0.21	0.39
Yu et al. (1979)	optimality of treatment for meningitis	HR	30	65

Table 2
Mean Difference of Transformed Effect Sizes by Type of Criterion

Criterion type	*N*	*M*	*SD*
Educational	18	0.09	0.96
Financial	5	0.20	1.53
Forensic	10	0.89	2.16
Medical	51	0.82	3.05
Clinical–Personality	41	0.19	4.83
Other	11	0.14	1.34

Note. All statistics are computed on weighted observations, with weights as explained in the text. $F(5, 130) = 2.11$, $p < .07$.

Educational, Financial and Clinical-Personality had smaller gaps between experts and algorithms, with a gap of 0.09, 0.20, and 0.19 standard deviations respectively. For forensic and medical predictions, the gaps were 0.89 and 0.82 standard deviations, respectively.

Moreover, the authors note, in instances where clinical experts had access to the statistical prediction rule, the statistical prediction rules still won. When the clinical experts had access to more data than the statistical prediction rule was using, the statistical prediction rule still won on average.

Even when the experts had access to both the statistical prediction rule, and more information than the SPR used, the SPR *still* beat out the experts.

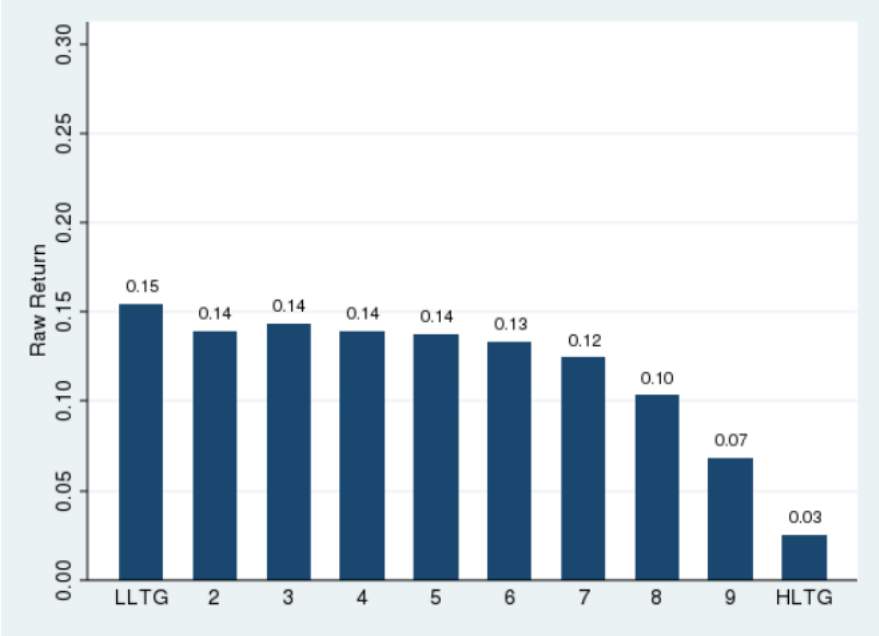
Now these statistical prediction rules were of course developed by experts themselves, but the fact is a layman with access to these statistical prediction rules would be able to out-predict a clinical expert in these fields most of the time.

This was from the year 2000, and so in all likelihood the gap is larger today.

2.3 Financial Predictions

From the paper "Diagnostic Expectations and Stock Returns"[2], the researchers looked at how well financial analysts predicted stock portfolio returned from the years 1981 to 2015.

[2]https://didattica.unibocconi.it/mypage/upload/154156_20170919_051604_DIAG_STOCKS_BGLS_NBER.PDF



They were ordered based on what the financial analysts predicted was high long-term growth and low long-term growth, with the stocks predicted as having the lowest rate of return having the highest rate of return, and stocks with the highest rate of return having the lowest rate of return.

Now when one thinks of academic credibility, they're usually not thinking about financial analysts. But financial analysts are a good litmus test because, once they get their degrees, they then have to go out into the real world and perform. And they perform very poorly. Significantly worse than random guessing.

The old adage of a dart throwing monkey being as good as a financial analyst is simply false. The dart-throwing monkey would radically outperform the financial analysts, and the cope that this is only true over the short term is false as well. As this paper looked at a random sample of financial analysts from 1981 to 2015.

Part of the reason financial analysts are taken less seriously than other academics is that, unlike other academics, they have to actually get out and make hard predictions with unambiguous outcomes. A sociology professor does not. In fact, even a biologist doesn't have to make these kinds of predictions.

2.4 Behavioral Economists vs. Mechanical Turk

Amazon has a program called "The Mechanical Turk," which links up people, usually poor people from poor country, who can be paid to do a long series of monotonous tasks on a computer - filling out forms, lists, tables.

A behavioral economist is a kind of psychologist who predicts how people will behave in economic contexts - what they will buy, where they'll try to work, what they'll invest in.

The paper "Predicting Experimental Results: Who Knows What"[3] by Vigna and Pope compared how a sample of opportunity (people they could get to participate in this study) predicted

[3]https://www.nber.org/papers/w22566.pdf

how much effort different groups of people were willing to put into various tasks. They compared PhDs in behavioral economics, PhD students, undergraduates and MBA students, and Mechanical Turk workers. The Mechanical Turk workers are anonymous but a 2009 paper showed that, at least 11 years ago, 57% of them were from the US, 32% from India, and the rest from Romania, Pakistan, UK, the Philippines and Canada.

Table 3. Accuracy of Forecasts by Group of Forecasters versus Random Guesses

	Average Accuracy (and s.d.) of *Individual* Forecasts	Accuracy of *Mean* Forecast (Wisdom of Crowds)	% Forecasters Doing Better Than Mean Forecast	Wisdom of Crowds: Accuracy Using Average of Simulated Group of Forecasters, Mean (and s.d.)	
				Group of 5	Group of 20
	(1)	(2)	(3)	(4)	(5)
Panel A. Mean Absolute Error					
Groups					
Academic Experts (N=208)	169.42 (56.11)	93.48	4.33	113.98 (23.15)	98.80 (11.68)
PhD Students (N=147)	171.42 (76.05)	91.65	8.16	117.99 (31.07)	97.78 (14.43)
Undergraduates (N=158)	187.84 (85.97)	87.86	3.16	115.46 (35.30)	94.80 (17.80)
MBA Students (N=160)	198.17 (86.04)	100.72	8.11	129.31 (34.34)	110.65 (17.05)
Mturk Workers (N=762)	271.57 (144.81)	146.93	17.85	173.01 (68.21)	150.93 (39.57)
Benchmark for Comparison					
Random Guess in 1000-2500	415.99				
Random Guess in 1500-2200	224.63				
Panel B. Mean Squared Error					
Groups					
Academic Experts (N=208)	49822 (34087)	12606	2.88	20046 (7894)	14438 (3234)
PhD Students (N=147)	53081 (50081)	11980	6.12	21365 (11268)	13895 (4142)
Undergraduates (N=158)	60271 (61112)	9769	2.53	19883 (12267)	12336 (4645)
MBA Students (N=160)	69855 (63213)	13334	3.90	24676 (12661)	16156 (4781)
Mturk Workers (N=762)	128801 (130473)	23660	9.71	44747 (32929)	28931 (13868)
Benchmark for Comparison					
Random Guess in 1000-2500	249534				
Random Guess in 1500-2200	75423				
Panel C. Rank-Order Correlation Between Actual Effort and Forecasts					
Groups					
Academic Experts (N=208)	0.42 (0.32)	0.83	4.81	0.65 (0.18)	0.76 (0.09)
PhD Students (N=147)	0.48 (0.30)	0.86	6.80	0.70 (0.18)	0.80 (0.09)
Undergraduates (N=158)	0.45 (0.31)	0.87	5.06	0.69 (0.17)	0.80 (0.09)
MBA Students (N=160)	0.37 (0.33)	0.71	18.52	0.56 (0.21)	0.67 (0.11)
Mturk Workers (N=762)	0.42 (0.35)	0.95	0.26	0.69 (0.20)	0.87 (0.07)
Benchmark for Comparison					
Random Guess in 1000-2500	0.00				
Random Guess in 1500-2200	0.00				
Panel D. Correlation Between Actual Effort and Forecasts					
Groups					
Academic Experts (N=208)	0.45 (0.29)	0.77	9.41	0.64 (0.16)	0.73 (0.09)
PhD Students (N=147)	0.51 (0.28)	0.86	4.86	0.72 (0.15)	0.82 (0.07)
Undergraduates (N=158)	0.49 (0.30)	0.89	3.90	0.72 (0.16)	0.84 (0.07)
MBA Students (N=160)	0.42 (0.32)	0.77	15.11	0.62 (0.19)	0.72 (0.09)
Mturk Workers (N=762)	0.43 (0.35)	0.95	0.00	0.70 (0.19)	0.88 (0.06)
Benchmark for Comparison					
Random Guess in 1000-2500	0.00				
Random Guess in 1500-2200	0.00				

In terms of the average accuracy of the forecasts, the Mechanical Turks were substantially worse with a higher mean error. However, when all taken together, the overall correlation between mechanical turk workers assessment and actual effort of the target was 0.43. Compared to 0.51 for

PhD students, 0.49 for undergraduates, 0.45 for academic experts, and 0.42 for MBA students.

Consider also that some proportion of Mechanical Turk workers will just click buttons to get their 10 cents and not bother trying to make accurate predictions. A random guy making a prediction about something is at least a volunteer - he's volunteering his prediction, he's putting some effort into it. The same cannot be said for every individual Mechanical Turk worker. And so this comparison of behavioral economists to Mechanical Turks should be taken with a grain of salt; i.e. the Mechanical Turks are probably less accurate than highly motivated amateurs.

One could respond that the academics and academics-in-waiting these Mechanical Turks are being compared to could also lack motivation in their predictions or their speciality may not be geared toward making predictions about individual effort in tasks, well then that makes the point just as well; which is that the academic training itself is not the relevant factor, but the individual person's knowledge.

Now lets be clear: the academics did outperform the mechanical turks slightly. But not to such a degree that you should take the credential of behavioral economist as the last word on ability to predict economic behavior. The PhD students only had a .08 greater correlation with the outcome here than the Mechanical Turk workers. Now if instead of mechanical turk workers you had highly motivated amateurs, and instead of an opportunity sample of PhD behavioral economics students you had PhD economic students who set out specifically to study this problem voluntarily and without prompting, the accuracy of both groups might be higher. But certainly the results of this experiment shouldn't make one leap to the conclusion that PhD students (the best predicting group) intrinsically motivated about a topic would make better predictions than a highly motivated amateur on that topic. And it seems intuitive that the gap between highly motivated amateurs and mechanical turk workers is greater than the gap between an opportunity sample of PhD students and PhD students intrinsically motivated about the topic.

2.5 Law Cases

The paper "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking"[4] looked at 83 Legal Experts and had them predict rulings of 171 supreme court cases in 2002.

The experts correctly predicted 59.1% of case outcomes, getting 40.9% of them wrong. When the outcome was unanimous, the experts correctly predicted which way the Supreme Court would rule 65.3% of the time.

A statistical prediction rule correctly predicted case outcomes based on a limited set of coded input variables 75.0% correctly overall, and 74.2% of the time when the Supreme Court ruled unanimously.

From the paper "Insightful of Wishful: Lawyers' Ability to Predict Case Outcomes"[5], 481 lawyers were asked before a case to state their minimum goal, and what their confidence was in achieving their minimum goal, and then the authors looks at the relation between a lawyer's confidence in achieving this minimum goal and the whether or not they did. The results were that what the lawyers predicted was very weakly related to the actual case outcomes, and that there was no difference in the accuracy of prediction between highly experienced and less experienced lawyers:

[4]https://www.researchgate.net/publication/241795963_The_Supreme_Court_Forecasting_Project_Legal_and_Political_Science_Appro
[5]https://sci-hub.tw/10.1037/a0019060

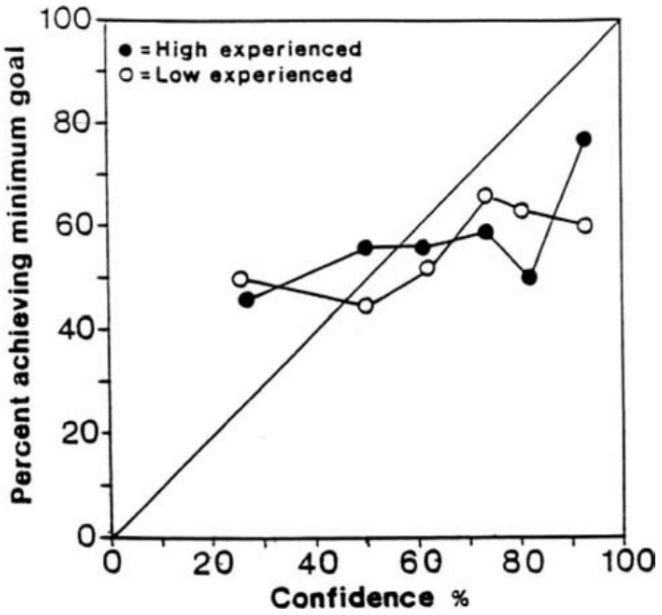


Figure 4. Calibration curves for lawyers with more than 10 years of experience (n = 231) and lawyers with 10 or fewer years of experience (n = 220).

Females were also less likely to overpredict their outcomes. In fact, below 80% confidence, females underpredicted their ability to achieve their minimum goal:

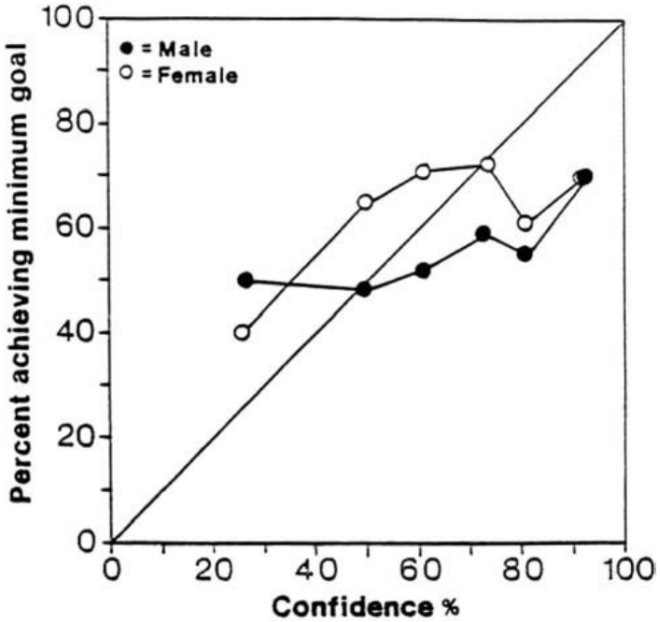


Figure 3. Calibration curves for male lawyers (n = 382) and female lawyers (n = 99).

In addition, it didn't seem to matter whether the lawyers were able to give a reason for their confidence. In fact, lawyers who were unable to give a reason for their level of confidence actually had slightly better outcomes than those who were able to give a reason:

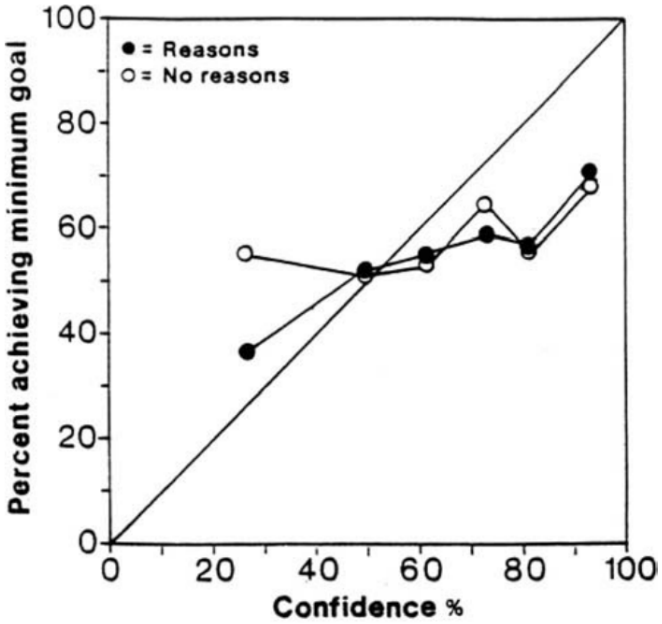


Figure 6. Calibration curves for the reasons group (n = 212) and the no-reasons group (n = 269).

The main takeaway is that the relation between lawyers' predictions and case outcomes is very close to a coinflip. Simply guessing every case has a 50-50 chance would not be significantly worse than how these lawyers predicted.

There is an important distinction to make here, and it can be made for both lawyers and doctors. A lawyer may be only slightly better than a coin toss at predicting case outcomes, and perhaps not actually do any better than a layman at predicting case outcomes. However, it is not then wise to reject legal counsel and argue for yourself in a trial. Similarly, a doctor may wish to operate on your knee, or your heart, in a way that does not benefit you and perhaps even does minor harm. This does not mean you should have your friend Bob do surgery on you.

Okay, if that's not our point, what then is our point with this? It's just to be realistic about what experts can and can't do. In terms of performing some kind of procedure, you want the medical professional to do it. But in terms of what the long-term benefits of a procedure will be, especially for a long-term ailment, the medical professionals will become more fallible.

Now the more acute the problem is, say a dislocated shoulder, the more medicine comes to resemble a hard science, and the more we should expect it to behave like a hard science. But the more long-term and ambiguous a condition is, the more we should expect medicine to behave like a soft science. And the more it behaves like a soft science, the smaller the gap between experts and laymen.

For lawyers predicting a case, that is more ambiguous, subjective and long-term. For merely arguing a case, that is a series of immediate decisions informed by training. Or for medicine,

perhaps heart stents actually have no long-term benefit for patients. But that is very different from saying a layman would be just as good at performing the stent surgery as a trained surgeon.

3 Not Supermen

3.1 IQ by Education Level

So aside from predictive ability, another thing we can look at is the IQs of people at various credential levels. This is typically not what people care about or why they consider someone to be an expert about something. Anecdotally, the man with the highest recorded IQ in the world, Christopher Langan, was a horse rancher from Bozeman Montana.

But it is a base we should cover.

There are several estimates for IQ by educational attainment that we have been able to find.

The National Longitudinal Survey of Youth tracked, from 1997 onward, 8,984 men and women born between 1980 and 1984, and tracked all manner of data on them from employment, income, criminality, family, and, of interest here, IQ, SAT and education level.

In 2011, The NLSY published data on credential level, and how this compared to a previous score on either the ASVAB - which is essentially a military IQ test - and the SAT.

Credential Level	ASVAB	SAT (M+V)	IQ derived from SAT
Professional (MD,DDS,JD)	121.15	1229.40	123.42
PhD	120.53	1265.54	126.34
Masters	111.30	1078.59	111.20
Bachelors	107.31	1039.95	108.07
Associate	99.54	940.25	100.00
HS Graduate	94.57	880.42	95.15

In 2014, the General Social Survey (GSS) recorded scores on a verbal IQ test, highest degree level, race, and hundreds of other factors. In 2014 Ryan Faulk and Sean Last cross-tabulated race, verbal IQ score and highest degree:

Verbal IQ by Race and Highest Degree Earned 1972-2014			
Highest Degree	**White Verbal IQ**	**Black Verbal IQ**	**Black/White Verbal IQ Gap**
High school Dropout	89	82	7
High school Diploma	98	90	8
Junior College Degree	102	95	7
Bachelors Degree	108	100	8
Graduate Degree	113	102	11

The white numbers compare similarly to the overall NLSY-97 results. Masters and above are collapsed into "graduate degree" for the GSS data. And of course the black verbal IQ is lower than the NLSY-97 results in all categories.

They also took scores from the National Adult Literacy Survey, which gave standardized tests on basic writing skills, the ability to read documents and common real-world math skills.

Here are the descriptions of the skills measured in the adult literacy survey:

> *"Prose literacy — the knowledge and skills needed to understand and use information from texts that include editorials, news stories, poems, and fiction; for example, finding a piece of information in a newspaper article, interpreting instructions from a warranty, inferring a theme from a poem, or contrasting views expressed in an editorial.*
>
> *Document literacy — the knowledge and skills required to locate and use information contained in materials that include job applications, payroll forms, transportation schedules, maps, tables, and graphs; for example, locating a particular intersection on a street map, using a schedule to choose the appropriate bus, or entering information on an application form.*
>
> *Quantitative literacy — the knowledge and skills required to apply arithmetic operations, either alone or sequentially, using numbers embedded in printed materials; for example, balancing a checkbook, figuring out a tip, completing an order form, or determining the amount of interest from a loan advertisement."*

Scores on these tests had a mean, median, and standard deviations. Which means Faulk and Last were able to convert them into "Adult Literacy IQ scores." Note that this is NOT a formal IQ test, but a skills test with scores presented the same way IQ scores are for the sake of comparison:

Estimated IQ by Race and Education			Black/White Gap (SD) by Literacy Type and Education			
Highest Degree	**White IQ**	**Black IQ**	**Document**	**Prose**	**Quantitative**	**Average**
High School Dropout	87	77	.59	.60	.79	.66
High School Graduate	99	90	.59	.57	.73	.63
2 Year Degree	104	94	.69	.60	.73	.67
4 Year Degree	113	103	.67	.65	.78	.70
Graduate Degree	121	110	.74	.70	.84	.76

These results are similar to the verbal IQ, ASVAB and SAT results from the NLSY-97 and GSS, though with generally a greater disparity between credential levels. The racial breakdowns are due to the focus of Faulk and Last's article being race differences in IQ when education level is held constant.

There are other lists floating around the internet, typically showing much higher scores than this. However we have not been able to find the original sources for those lists or what methods they used. But since this is an ongoing project, if you can find any additional studies we will add them to this section.

The purpose of this IQ data is simply to say that highly credentialed academics are not supermen in terms of intelligence as inferred from IQ tests. While an IQ of 120 puts one above the 90th percentile, it is not some unreachable level of intelligence for even the average person, let alone a highly motivated amateur with an IQ above this.

If you have tried to read academic papers, they are filled with all manner of jargon and terms that appear incomprehensible. However, any piece of jargon can be understood by anyone. There is nothing in an academic paper that is beyond your ability to understand. Now understanding enough of the terminology, and putting in the time and effort to be able to do so may not be something you'll ever do, but just know that the complexity comes from the interaction of simple elements, all of which you can understand. And either with a great deal of time reading papers to where these terms become second nature, or by putting in a great deal of effort to just read one paper - there is nothing fundamentally impenetrable about academic papers even to someone of average or slightly below-average intelligence.

The consensus of scientific fields is formed by people who, on average, have IQs around 123. Perhaps lower for social sciences. They are not some exalted beings, but real people who we would say are "pretty smart" but not immune to the foibles of man.

3.2 Knowledge Decay

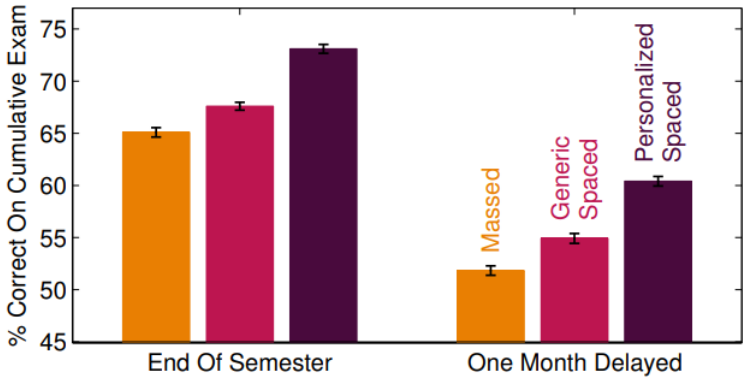
But they're not experts for their IQ, they're experts for their knowledge. Well lets look into that. Before someone becomes a research scientist, they are first a student.

There are several studies on academic and student knowledge retention. And knowledge retention of course varies based on how much you use it. From the paper "Improving students long-term knowledge retention through personalized review"[6] they looked at the knowledge decay of Spanish students at the University of Colorado.

And they found that over the 28 day intersemester break there was a major decline. The paper claims there was an 18.1%, 17.1% and a 15.7% decline in the personalized spaced, generic spaced and massed review conditions.

The paper was on review methods to improve knowledge retention.

However, upon looking at the values ourselves, we are unaware of where the authors got their decline numbers from - or if they are referring to an absolute or relative decline in scores. Here's the chart presented by Lindsey et. al.:



And here are those resulted tabulated, along with the absolute and relative declines for the three review groups:

Review Group	% Correct on Final Exam	% Correct on one month follow-up	Absolute Decline	Relative Decline
Personalized Spaced	73.1	60.7	12.4	17.0
Generic Spaced	67.6	55.2	12.4	18.3
Massed	65.2	52.2	13.0	20.0

[6] https://www.cs.colorado.edu/ mozer/Research/Selected%20Publications/reprints/LindseyShroyerPashlerMozer2014.pdf

They show absolute declines of 12.4%, 12.4% and 13.0%, and 17.0%, 18.3% and 20.0% for the Personalized Spaced, Generic Spaced and Massed review groups respectively.

Now assuming the raw data itself in this study isn't suspect, in this sample we saw an approximately 18.4% relative decline in one month. The question is - does this decline continue logarithmically - declining 18.4% off of the previous month's level until it reaches virtually nothing? Perhaps there is some rock bottom where knowledge isn't lost. For example, 10 years after learning Spanish, perhaps you remember how to count to ten and that "por que" means "why."

The paper "How Fast Do Students Forget What They Learn in Consumer Behavior? A Longitudinal Study"[7] spent 4 years retesting 374 students who took a consumer behavior class at an unnamed "Western University."

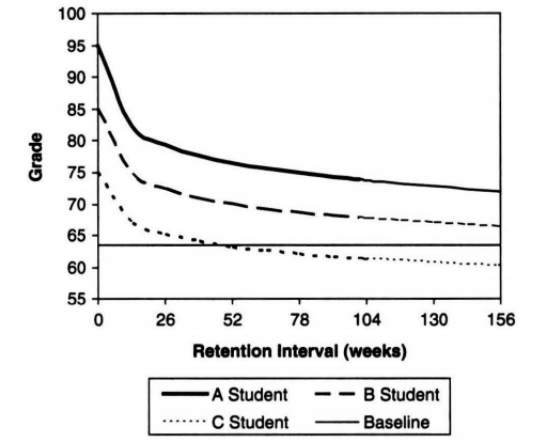


FIGURE 1: Expected Grade by Retention Interval for Hypothetical Students
NOTE: Lines are shown thinner after 101 weeks to reflect projections beyond the data in hand.

The results are displayed on the graph. Of note is that C-students were found to perform below "baseline" after a little less than a year out. "Baseline" being the score achieved by a random sample of students who never took the course but were just handed the final. They scored around 64%.

The important takeaway is that this decline occurred among A-students as well and was just as severe. Knowledge decay - at least within the range of these students - was not less at the higher abilities. So there is no reason to assume that if you go higher than this - once you get into the PhDs - that this trend won't continue up to that level. The knowledge decline could be even more severe since there is more knowledge to lose, or it could be less since perhaps they use it more, or maybe those things cancel out.

From "How Much Do Students Remember Over Time? Longitudinal Knowledge Retention in Traditional versus Accelerated Learning Environments"[8], the authors looked at how 270 first-year and fourth-year epidemiology students at various Canadian Universities retained knowledge in either traditional or compressed "supercourses."

[7] http://www.uky.edu/ kdbrad2/EPE773R/StudentPapers/ConsumerBehavior.pdf

[8] http://www.heqco.ca/SiteCollectionDocuments/How%20Much%20Do%20Students%20Remember%20Over%20Time.pdf

Table 4-3: Success of retention quizzes at each time point

	Baseline	3 months	6 months	12 months
T1				
Repeat	15 (100)	9.25 (61.6)	7.12 (47.5)	5.86 (39.1)
Non-repeat	5 (100)	3.07 (61.4)	2.61 (52.2)	2.39 (47.8)
S1				
Repeat	15 (100)	8.90 (59.3)	6.60 (44.0)	5.25 (35.0)
Non-repeat	5 (100)	2.55 (51.0)	2.40 (48.0)	2.55 (51.0)
T4				
Repeat	15 (100)	8.94 (59.6)	6.29 (41.93)	4.89 (32.6)
Non-repeat	5 (100)	3.09 (61.80)	2.86 (57.2)	2.29 (45.8)
S4				
Repeat	15 (100)	8.32 (55.46)	5.64 (37.6)	4.36 (29.1)
Non-repeat	5 (100)	2.54 (50.8)	2.71 (54.2)	2.36 (47.2)

Note: Values in brackets indicate the success of retention quizzes expressed as a percentage

The students were in four categories: Year 1 course in the traditional format, year 1 in the supercourse format, year 4 course in the traditional format, year 4 course in the supercourse format. All students were then tested 3 months, 6 months and 1 year after taking the course.

This and the Colorado study were focusing on how student retention varies by course type and review type, but we aren't particularly interested in that since we're just looking at the big picture of how much academics retain. And all academics, before they are academics, they are students.

On the repeat questions, after 1 year the scores for year 1 and year 4 students fell 60.9% and 70.9% respectively.

On the non-repeat questions, scores only fell between 49% and 54.2%.

The important point from this study is that Year 4 students - in both the traditional and supercourse setting - had slightly more knowledge decay than the year 1 students in both the supercourse and traditional course settings.

Now this is just one study, and the sample size is small enough that this could just be an artifact. However, you certainly cannot say that at the higher levels there is less knowledge decay over the same amount of time. There appears to be roughly as much, with weak evidence that there is slightly more knowledge decay at higher levels.

From the paper "Knowledge loss of medical students on first year basic science courses at the university of Saskatchewan"[9], showed higher retention rates than the previous study.

[9]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1397826/

Table 1

Comparing the scores on tests and re-tests of knowledge for three basic science courses

Course	Exam %	Test % on selected questions	Re-test % on selected questions	Relative Knowledge Loss[1]	Course Evaluation	Correlation between Test re-test scores
Neuroanatomy N = 24	82.5	87.7	41.5	52.7	3.6/6.0 (60%)	.310 p = .140
Immunology N = 29	77.0	74.8	61.7	17.6	4.1/6.0 (68%)	.619 p < .001
Physiology N = 25	83.2	Not available	67.1	19.4[2]	4.5/6.0 (75%)	.523 p = .007

They showed declines of only 17.6% and 19.4% for physiology and immunology students, but 52.7% for Neuroanatomy students.

These knowledge retention studies have been focused on retention at one year or less, which is understandable given how difficult it would be to retest people 10 years out. However, there is one study on economics students that went out much longer than these.

The paper, entitled "Factors Determining Student Retention of Economic Knowledge after Completing the Principles-of-Microeconomics Course"[10], and they retested 59 economics students in 1979, who had completed the course between 1971 and 1976. So between 3 and 8 years on.

The results weren't presented in an easy to understand table or chart, but they found that knowledge in these students decayed at roughly 18% per year, and this was a continuous logarithmic trend.

So after 1 year you only know 82% of what you knew. At year 2 you're at 67.24%, at year 3 you're at 55.14%, and by year 8 you're at 20.44%. So at least in this sample of economics students at that time, the knowledge decay continued logarithmically.

How applicable is this to other fields today? Well, would knowledge decay faster or slower in the past? Intuitively it seems it would decay slower in the past since there was less information overload as there was less television and no internet.

What about other disciplines - would they decay more or less than economics? There are reasons to suspect more, because economics is, whatever your opinions are of it as a serious discipline, it is a highly integral discipline, which is to say economic concepts aren't isolated from each other. This is just speculation, which is confirmed by the previous data presented on other fields from more recent studies which show higher decay rates than for these economics students. But confirmed speculation is by no means conclusive.

It just means that, for now, there's no reason to think that this study, from 1979, doesn't apply to people today. If anything knowledge decay greater in fields outside of economics, and greater today than in the past.

Knowledge retention is something we really should know more about. It's quite profound how little research there is, at least that we have been able to find, given how many billions of dollars

[10] https://sci-hub.tw/10.2307/1182376

and millions of life-years are pumped into college.

3.3 Bullying

Now there a a lot of studies on the prevalence of bullying in academia. Unfortunately, their definitions and methods vary wildly, as do their timescales. And it doesn't tell us how bullying among academics compares to bullying among the general population. The paper "Faculty Experiences with Bullying in Higher Education" looks at 11 studies on the matter, and the average reported rate of having been bullied was 32.83%. However both the definitions and timescales varied with each study, and there's no private-sector population to compare them to.

The book "Bullying and Harassment in the Workplace"[11] had a table of 124 studies on workplace bullying prevalence among various organizations and populations, but the authors also coded the definitions of bullying used.

Here's what their table looked like:

TABLE 3.3

Studies on the Frequency of Workplace Bullying

Country	Authors	Sample	No.	Definition*	Prevalence
Austria	Niedl (1995)	Hospital employees	368	1b + 3a	26.6% in sample; 7.8% of the population
		Research institute employees	63	1b + 3a	17.5% in sample; 4.4% of the population
Belgium	Notelaers and De Witte (2003)	Association of local government, consulting office, nonprofit organisation, print office, chemical production	873	8	16%
Belgium	Notelaers et al. (2006)	18 organisations	6175	1a + 3a 7	20.6% 3.1%
Croatia	Russo et al. (2008)	Schoolteachers	764	1b + 3b	22.4%
Denmark	Hogh and Dofradottir (2001)	Randomised sample	1857	5	2%
	Mikkelsen and Einarsen (2001)	Course participants at the Royal Danish School of Educational Studies	99	1b + 3a + 4	4: 2%; 1b 3a: 14% (7.8% for a more stringent criterion)
		Hospital employees	236	1b + 3a + 4	4: 3% now and then; 1b 3a: 16% (2%)
		Manufacturing company	224	1b + 3a + 4	4: 4.1% now and then; 1b 3a: 8% (2.7%)
		Department store	215	1a + 3a + 4	4: 0.9%; 1b 3a: 25% (6.5%)
	Mikkelsen and Einarsen (2002)	Danish manufacturing company	224	1a + 3a + 6a 1a + 3a + 6b	8% 2.7%
	Agervold and Mikkelsen (2004)	Danish manufacturing company	186	1a + 3a + 6a 3a + 4 3b + 4	13% 1.6% 10.3%

[11] https://libgen.lc/ads.php?md5=82DAF5926DD30AB54238708BEDBFA851

	Hubert and van Veldhoven (2001)	Sample including a variety of branches	66764	2 + 5	2.2% mean of 4 items referring to aggressive and unpleasant situations often or always
Norway	Matthiesen et al. (1989)	Nurses and assistant nurses	99	1a + 4	3.9%
		Teachers	84	1a + 4	10.3%
	Einarsen and Skogstad (1996)	14 different samples; total	7787	1a + 4	Weekly 1.2% (yes, now and then: 3.4%)
		Health and welfare managers	344		8.6% occasional bullying
		Psychologists' union	1402		0.3% (12.0%)
		Employers' federation	181		0.6% (2.3%)
		University	1470		0.6% (2.3%)
		Electricians' union	480		0.7% (2.8%)
		Health-care workers	2145		0.8% (3.1%)
		Industrial workers	485		1.1% (2.2%)
		Graphical workers' union	159		1.3% (6.5%)
		Teachers' union	554		1.9% (8.9%)
		Trade and commerce	383		2.4% (2.0%)
		Union of hotel and restaurant workers	172		2.9% (4.3%)
		Clerical workers and officials	265		2.9% (4.1%)
	Einarsen et al. (1998)	Representative sample from a county	745	1a + 4	3%. 8.4% with previous experience
	Eriksen and Einarsen (2004)	Nurses	6485	3a + 4	4.5%
	Hauge et al. (2007)	General working population	2539	1a + 3a	1.9%
	Matthiesen and Einarsen (2007)	Six Norwegian labour unions	4742	1a + 4	8.3%
	Glaso et al. (2009)	General working population	2539	1a +3a + 4	4.1%
	Magerøy et al. (2009)	Royal Norwegian Navy	1604	1a + 4	2.5%
					(*continued*)
Sweden	Leymann (1992)	Handicapped employees; nonprofit organisation	179	1b + 3a	8.4%; 21.6% handicapped; 4.4% not handicapped
	Leymann and Tallgren (1993)	Steelworks employees	171	1b + 3a	3.5% (probably lower because of dropouts)
	Leymann (1993a)	Sawing factory	120	1b + 3a	1.7%
	Leymann et al. in Leymann (1993b)	Nursery schools	37	1b + 3a	16.2%
	Leymann (1993a, 1993b)	Representative of employed except self-employed	2438	1b + 3a	3.5%
	Lindroth and Leymann (1993)	Nursery school teachers	230	1b + 3a	6%
	Hansen et al. (2006)	Pharmaceutical	91	1a + 4	2%
		Telecommunication	101	1a + 4	5%
		High school	172	1a + 4	7%
		Wood industry	34	1a + 4	6%
		Social insurance	39	1a + 4	3%
Turkey	Cemaloglu (2007)	Schoolteachers	337	1a + 3b	6.4%
	Soylu et al. (2008)	General working population	152	1a + 3a	48%
	Ozturk et al. (2008)	Academic nurses	162	1c + 3b	20.4%
	Yildirim et al. (2007)	University nursing school academics	210	1b	17%
	Yildirim and Yildirim (2007)	Nurses from the European side of Istanbul Province	505	1c + 3b	86.5%
	Bilgel et al. (2006)	Public-sector organisations	877	1b + 3b	55%
United Kingdom	Rayner (1997)	Part-time students	581	1c + 4	53%
	UNISON (1997)	Public-sector union members	736	1 + 4	14%; 1c+4: 50%
	Quine (1999)	National Health Service	1100	3b	38% persistently bullied within last 12 months
	Cowie et al. (2000)	International organisation	386	4	15.4%
					(*continued*)

Going through this, we found 6 studies on bullying from this table which looked at an academic population for which also existed at least one other study on bullying outside of academia using the same definition. So we have a few apples to apples comparisons.

The 1b + 3a definitions:

--- 1b + 3a Definition studies				
Niedle	1995	Austria	Hospital Employees	7.8%
			Research Institute Employees	4.4%
Mikkelsen & Einarsen	2001	Denmark	-Students at Royal Danish School of Ed. Studies	14%
			Hopsital Employees	16%
			Manufacturing Company	8%
			Department Store	25%
Minkel	1996	Germany	Rehab clinic Admin.	8.7%
Mackensen	2000	Germany	"Administration"	2.9%
Muhlen	2001	Germany	"Communal Admin."	10%
			Military Admin.	10.8%
Kaucsek	1995	Hungary	Army	5.6%
			Bank Employees	4.9%
			Bank inspectors	2.5%
Leymann	1992	Sweden	Employees at nonprofit	4.4%
			(Handicapped)	21.6%
Leymann	1993	Sweden	Steelworkers	3.5%
Leymann	1993	Sweden	Sawmill	1.7%
Leymann	1993	Sweden	Nursery School	16.2%
Leymann	1993	Sweden	Sample of workers	3.5%
Lindroth	1993	Sweden	Nursery School Teachers	6%
Cemaloglu	2007	Turkey	Schoolteachers	6.4%

The 3b definition studies:

--- 3b Definition Studies			
Gil-Monte 2006	Spain	Employees working with Disabled people	19%
Justicia 2007	Spain	University Staff	9%
Quine 1999	Britain	National Health Service	38%

The 1a + 4 definition studies:

--- 1a + 4 Definition Studies

Ortega 2008	Denmark	Nursing Homes	1.6%
Vartia 2002	Finland	Prison Officers	11.8%
Matthiesen 1989	Norway	Nurses	3.9%
Einarsen 1996	Norway	Health/Welfare manag.	8.6%
		Psychologists Union	12%
		Employers' Federation	2.3%
		University	2.3%
		Electrician's Union	2.8%
		Healthcare workers	3.1%
		Industrial Workers	2.2%
		Graphical Workers	6.5%
		Teachers' Union	2.0%
		Trade and Commerce	2.4%
		Hotel and Restaurant	4.3%
		Clerical Workers	4.1%
Einarsen 1998	Norway	General Population	3%
Einarsen 2007	Norway	6 Labor Unions	8.3%
Mageroy 2009	Norway	Navy	2.5%
Nielsen 2009	Norway	General Workforce	4.6%
Hansen 2006	Sweden	Pharmaceutical	2%
		Telecom	5%
		High School	7%
		Wood industry	6%
		Social Insurance	3%
Coyne 2004	Britain	Firefighters	16%

The 1b definition studies:

--- 1b Definition Studies

Merecz 2006	Poland	Nursing Staff	69.6%
Fornes 2008	Spain	Professional School Nurses	17.2%
Yildirim 2007	Turkey	University nursing school academics	17%

And the definition 4 studies:

```
-- 4 Definition studies

Vartia 1996             Finland         Municipal Employees     10.1%
Kivimaki 1996           Finland         Hospital Staff          5.3%
Piirainen 2000          Finland         General Workforce       4.3%
Kivimaki 2004           Finland         Hospital Employees      4.8%
Meschkutat 2002         Germany         General Workforce       5.5%
O'Moore 2000            Ireland         Random Sample           16.9%
Hubert  2001            Holland         "Office Businesses"     4.4%
Justicia 2007           Spain           -University Employees   24.1%
Escartin 2008           Spain           General Workforce       10%
Cowie 200               Britain         "International Org."    15.4%
Jennifer 2003           Britain         "3 large orgs"          21.1%
Paice   2004            Britain         21 Hospitals            18%
Quine 2002              Britain         Junior Doctors          37%
Coyne 2003              Britain         Public Sector Org.      39.6%
```

Research institute employees came in at 16th out of 21 for the 1b + 3a definition studies. Students at the Danish Royal School of Educational Studies came in 5th out of 21. So in terms of their combined rank, they rank at 10.5, or precisely average in terms of rank-order.

The average for the whole of the 1b + 3a definition studies was 8.76%, whereas the two academic studies were 4.4% and 14%, averaging 9.2%. Now perhaps you don't think the bullying rate of students should reflect academics, and research institute employees are the only ones that really matter because these are the people actually writing the studies. But given the imprecision of the data, any data on tangentially-academic fields should be considered in a holistic sense. But if you think not, that's fine.

The research institute employees still finished 16th, not 21st on bullying rate. And their rate was 4.4%, not 0%. And it was higher than bank inspectors, sawmill workers, steel workers and a general sample of employed persons in Sweden.

Since the point of this is to not just say, but really show and drive home that academics are not wizards, so long as the point that they are not wizards is taken away, that's fine.

The 3b definition studies showed University employees having much lower bullying rates, but this was only compared to NHS workers at 38% and employees working with disabled people at 19%. It's difficult to argue much of anything can be drawn from the 3b definition studies comparison.

In the 1a + 4 definition studies, we have some gold: a single researcher using the same definition and survey questions across multiple industries with the data presented in the same study: Einarsen 1996. This showed University workers coming out a little better than average in terms of bullying rates. They finished tied for 9th/10th out of 12 in rank order, just above Teachers Unions and Industrial Workers, and just behind Trade and commerce at 2.4%. The average for the 12 Einarsen samples was 4.38%, but this average was dragged up by Health and Welfare managers and the Psychologists Union (who may be reporting high results because as psychologists they might look for and be more sensitive to bullyinh and find false positives, similar to how psychology students are more likely to think they have mental health problems). The median of the Einarsen studies is 3.6%. So again we see University employees being on the low end of bullying, but not at the bottom, and having a rate very close to industrial workers, trade and commerce and the electrician's union.

The 1b definition studies don't have enough meaningful comparisons to be drawn.

The definition 4 studies, which is on University employees in Spain, has a rather eye-poppingly

high result. Out of the 14 definition 4 studies, they rank the 3rd highest. Of course "University Employees" could include janitors, groundskeepers, etc. But even so, based on the data of industrial workers, blue-collar workers tend to have rather low bullying rates, and so do academics in all honesty. It's just one study - maybe it's a situation particular to Spain as a country, or that University in particular.

The paper "Destructive Conflict and Bullying at Work"[12] by Hoel and Cooper also looked at bullying rates across industries using a the same definition and questions across industries in the United Kingdom, so like Einarsen 1996, this is apples-to-apples gold.

Table 4: 'Current, past and 'indirect bullying'

Sector	Bullied last 6 months (%)	Bullied last 5 years (%)	Witnessed bullying last 5 years (%)
Post/Telecom.	16.2	27.9	50.4
Prison	16.2	32.1	64.0
Teaching	15.6	35.9	57.7
Other	14.3	20.0	40.0
Dance	14.1	29.6	50.0
Police Service	12.1	29.2	46.4
Banking	11.6	24.6	39.6
Voluntary Org.	10.7	26.7	55.6
NHS Trusts	10.6	25.2	47.2
Local Authority	10.5	21.4	42.7
Civil Service	**9.9**	**25.7**	**47.1**
Fire Service	8.9	20.0	43.2
Hotel industry	7.5	16.8	46.3
High. Educ.	7.2	21.3	42.8
Retailing	6.8	17.6	33.7
Manufacturing	4.1	19.2	39.0
Totals	10.6	24.7	46.5

Hoel and Cooper found higher education at 7.2% of the respondents in Higher Education saying they have been bullied (by their definition of bullying) in the past 6 months. The unweighted average of all of these categories is 11.05%. Hoel and Cooper may have done an n-weighted average, which seems inappropriate for this topic given that n-weighting doesn't necessarily reflect the employment distribution of the general population.

Either way, you see once again, universities being on the low end for bullying rate, but not at the very bottom, edged out by manufacturing and just between retailing and the hotel industry.

What is the purpose of this data on bullying? Is it because we care a lot about bullying in particular? Well obviously we're against it as are most people, but the point is to say that academics are not immune to human foibles. They are lower, no doubt, but they are not so low that you can say they are immune from base and petty impulses.

Now in theory, in the absence of any evidence on the matter, you shouldn't assume they are any more or less biased than the general population. But in practice, we deal with the problem of "magic science man." Where magic science man casts a spell, proclaiming for example "there are no innate sex differences in personality." And then the wizard cultists just believe it, and then if you try to argue against the wizard cultists against this, they will be impervious to reason. If they can't argue against you, they will appeal to the authority of the wizards - wizards who know the

[12]http://www.bollettinoadapt.it/old/files/document/19764Destructiveconfl.pdf

ways of the world and aren't biased.

The point of this was to show that wizards don't exist. Or at least they don't exist as a class; i.e. the certification of being a wizard doesn't make you a wizard.

3.4 On Psychology

Before going further into critiques on the capabilities of academics in general, we should talk about Psychology. Because many of these analyses of academic competence involve psychologists - and the viewer may be tempted to hand-wave away such information by imagining it to be some problem specific to psychology.

While there is no singular indicator of a field's "rigor," one indicator we can look at is statistical power and replication rate.

Statistical power is the probability that a statistical test will reject a false null hypothesis. Which means that it won't fail to find a significant effect when one exists in reality. It's a quantifiable way to tell if a study is even capable of finding a positive effect and rejecting the null hypothesis.

And there have been multiple field-wide meta-analyses of either randomized or most commonly cited studies in various fields, and the statistical power for those fields. Sean Last collected several of these studies[13], and looked at the statistical power for each field on the basis of these studies:

Citation	*Discipline*	*Mean / Median Power*
Button et al. (2013)	Neuroscience	21%
	Brain Imaging	8%
Smaldino and McElreath (2016)	Social and Behavioral Sciences	24%
Szucs and Ioannidis (2017)	Cognitive Neuroscience	14%
	Psychology	23%
	Medical	23%
Mallet et al (2017)	Breast Cancer	16%
	Glaucoma	11%
	Rheumatoid Arthritis	19%
	Alzheimer's	9%
	Epilepsy	24%
	MS	24%
	Parkinson's	27%
Lortie-Forgues and Inglis (2019)	Education	23%
Nuijten et al (2018)	Intelligence	49%
	Intelligence – Group Differences	57%

Now the general findings here are quite shocking. Of the fields and analyses Sean found, only intelligence and intelligence group differences (subsets of psychology) have anything around

[13]https://ideasanddata.wordpress.com/2020/06/25/on-trusting-academic-experts/

50% statistical power. And this point will be important later when we start talking about group differences in intelligence. But for now, the important point is that psychology is not a particularly weak field in terms of statistical power.

We can also look at replication rates. In July 2016 Nature did a survey of 1,576 of various fields and asked what percentage of finds they believed could be replicated. These were the results (Table from Last):

Discipline	Estimated Replication Rate
Physics	0.73
Other	0.52
Medicine	0.55
Material Science	0.60
Engineering	0.55
Earth and Environmental Science	0.58
Chemistry	0.65
Biology	0.59
Astronomy	0.65

In 2015, the Open Science Collaboration Project took 100 psychology experiments and attempted replications, and were only able to replicate 47 of them, or 47%. This was then used as a ringing condemnation of psychology in particular. But this is not much lower than the replication rates researchers think their fields would have in what the public might consider to be hard sciences.

Last also collected a series of replication analyses of various fields (not surveys):

Citation	Field	Replication Rate
Soto (2019)	Differential Psychology	87%
Cova et al. (2019)	Experimental Philosophy	70%
Camerer et al. (2016)	Economics	61%
OSC (2015)	Social Psychology	25%
OSC (2015)	Cognitive Psychology	50%
Prinz (2011)	Pharmacology	21%
Begeley and Ellis (2012)	Oncology	11%
Neuroskeptic (2014)	Neuroscience	6%

Different types of psychology are broken down on this list. Most relevant for this course in particular is differential psychology, which deals with how people differ on all manner of psychological traits. In this light, psychology in general does not appear to be particularly bad. Moreover, given the greater ambiguity of psychology and the greater difficulty in operationalizing psychological experiments than say a physics experiment, the fact that psychology has a lower replication rate than biology (which isn't, to our knowledge, even established) wouldn't necessarily mean psychologists

are less statistically literate than biologists; it could just reflect the more ambiguous nature of the subject matter.

The reason for making this point relatively defending psychology is that several of the studies we will present on statistical knowledge focus on psychologists. And the defense is relative to other fields, it is not an absolute defense. The reason for this defense is because several of the studies we will be presenting on general statistics knowledge focus on psychologists, and we are cutting the escape rope of "this only applies to psychologists," when there is no reason to presume other fields are better.

3.5 Knowledge of Statistics

Proper knowledge of statistics is vital in all forms of research. To understand anything about the world, you must understand it at a quantitative level. This is something anyone can do. Anyone can make a table, and distil much of the complexity of the world into a few indicative statistics.

In order to understand, for example, how wealthy Ireland is, you cannot just travel to Ireland and jot down your impressions. Which part of Ireland did you see? How many houses did you see? Do you have the ability to evaluate the value of goods, houses, industrial equipment, cars, clothes - just from looking at them? It is impossible to understand the world from a purely experiential level, and so you must use statistics.

Of course it is possible to lie with statistics. But it is much easier to lie with anecdotes. A statistic almost always has some source - government agency, a study - and while those can be fabricated, it is more difficult to do. Moreover, there are statistical plausibility tests which can be used on all manner of data which would draw red flags to statistics. While these tests may not catch fraud 100% of the time, they serve as an additional barrier to lying about the raw data itself.

However, most of the time, raw data isn't outright fabricated. The way people lie with statistics is either cherry-picking and hoping the viewer doesn't know about the cherries-not-picked, or come up with some inclusion criteria that excludes data which would otherwise be relevant to the point you're making.

It's all terribly complicated, and there's no simple way to determine if a statistical argument is correct or incorrect, or honest or dishonest. You just have to know things.

Unfortunately, even if they don't say so, the general public usually outsources the answer to these problems with "academic consensus." And there are several problems with this as stated before.

However, in this section we will just be dealing with the question of general statistical competence of academics - are they much better than that of highly motivated amateurs? Well, perhaps, as a rule, professional researchers have had more statistical training. And this fact alone - in the absence of any other information - would make it prudent to assume that they're generally better at interpreting statistics.

But we have more information. We have several studies which tested the statistical knowledge of academics, and what they found was that, in terms of dealing with the kinds of problems one faces when doing real-world experiments, academics come off rather poorly (Getting the equivalent of "D" and "F" grades on most of these tests) and not significantly better than first-year college students - which is relevant to the viewer who likely has the same level of education as a first year college students.

So lets get into it.

3.5.1 McShane 2016

The paper "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence"[14] by McShane and Gal gave several simple statistical questions to multiple types of academics to see how well they could answer them.

One of the questions - presented below - was presented to 75 researchers who had published in the New England Journal of Medicine. The question was as follows:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how different interventions might affect terminal cancer patients' survival. Participants were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Participants were then tracked until all had died. Participants in Group A lived, on average, 8.2 months post-diagnosis whereas participants in Group B lived, on average, 7.5 months post-diagnosis (p = 0027).

Group	Intervention	Life Expectancy
Group A	"Write daily about positive things you were blessed with"	8.2 months
Group B	"Write daily about misfortunes others had to endure"	7.5 months

> Which statement is the most accurate summary of the results?
>
> (a) Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was greater than that lived by the participants who were in Group B.
>
> (b) Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was less than that lived by the participants who were in Group B.
>
> (c) Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was no different than that lived by the participants who were in Group B.
>
> (d) Speaking only of the subjects who took part in this particular study, it cannot be determined whether the average number of post-diagnosis months lived by the

[14] https://sci-hub.tw/10.1287/mnsc.2015.2212

participants who were in Group A was greater/no different/less than that lived by the participants who were in Group B.

Very verbose. Here's a laymen's phrasing of the question: which group lived longer? Group A, Group B, they lived the same, or it cannot be determined? This is *not* a trick question. It is in fact as simple as it seems. Which group lived longer?

Yes, the correct answer is Group A.

Here were the results from the authors who had published in the *New England Journal of Medicine* by p-value assigned to the results:

Table 1 Study 1 Results

Option	$p = 0.01$	$p = 0.27$
	(a) Wording 1	
A	95	10
B	0	0
C	0	55
D	5	35
n	20	
	(b) Wording 2	
A	83	22
B	0	0
C	0	35
D	17	43
n	23	
	(c) Wording 3	
A	88	3
B	3	0
C	6	62
D	3	34
n	32	

Notes. Each cell gives either the percentage of participants who gave the given response option or the sample size. Participants are much more likely to correctly choose option A when $p = 0.01$. The response wording has no substantial impact on the results.

When the authors were presented the data with a p-value of 0.01, they answered "A" (the correct answer) 95% of the time, 83% of the time and 88% of the time. Pretty good. Perhaps not as high as you'd hope, but pretty good.

But when the p-value was 0.27, they answered "A" only 10%, 22% and 3% of the time.

What this shows is that the authors of articles published in the New England Journal of Medicine overwhelmingly failed to distinguish statistical significance from descriptive statistics. A p-value of 0.27 is not a statistically significant result, but these authors then went on to say what happened didn't actually happen because it wasn't statistically significant.

Say your vertical leap is 4 feet, and bob's vertical leap is 3 feet. Whose vertical leap is higher? Yours. Oh, but n=2? That's statistically insignificant, we cannot say you can actually jump higher than Bob given such a small sample! That's what they're doing here.

As a layman, this appears incomprehensibly stupid. However, there is a phenomenon among military aircraft pilots called "CFIT" or "Controlled Flight Into Terrain." It is where a pilot is so focused on his instruments and making sure the aircraft is running properly that he loses track of where the aircraft is headed relative to the earth - something you could see by just looking out

of the window. It's not that the pilot is incomprehensibly stupid, it's that he is so fixated on his instruments that he ends up losing track of the "big simple."

And that's what happened here: academics so fixated on their statistical instruments that they lose track of the big simple.

Similar questions were given within the paper. The second question was given to 299 researchers who had published in the *American Journal of Epidemiology*:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease.
>
> A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a p-value of 0.175.
>
> Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?

The answers the researchers could choose from can be paraphrased as:

(a) A random person taking drug A would be **more likely** to recover than someone taking drug B

(b) A random person taking drug A would be **less likely** to recover than someone taking drug B

(c) A random person taking drug A would be **equally likely** to recover than someone taking drug B

(d) **It cannot be determined**

Again, this is not a trick question. Even a statistically insignificant result does not mean the effect isn't real; just that the probability of the results being caused by something other than the difference in treatment, or just random noise, is higher. Like with the first question, the authors were given asked to answer the same question but with the p-value manipulated:

Table 2 **Study 2 Results**

	Small treatment difference				Large treatment difference			
Option	$p = 0.025$	$p = 0.075$	$p = 0.125$	$p = 0.175$	$p = 0.025$	$p = 0.075$	$p = 0.125$	$p = 0.175$
				(a) Judgment				
A	70	16	25	16	81	21	24	22
B	0	0	0	0	0	0	3	0
C	10	22	34	38	3	35	15	16
D	20	62	41	47	16	44	58	62
				(b) Choice				
A	87	50	53	41	94	53	52	49
B	0	0	0	0	0	0	0	0
C	13	50	47	59	6	47	48	51
n	30	32	32	32	31	34	33	37

Notes. Each cell gives either the percentage of participants who gave the given response option or the sample size. Participants are much more likely to choose option A for both the judgment question and the choice question when $p < 0.05$, and there is no substantial variation in the likelihood of choosing option A across the three $p > 0.05$ conditions. The magnitude of the treatment difference has no substantial impact on the results.

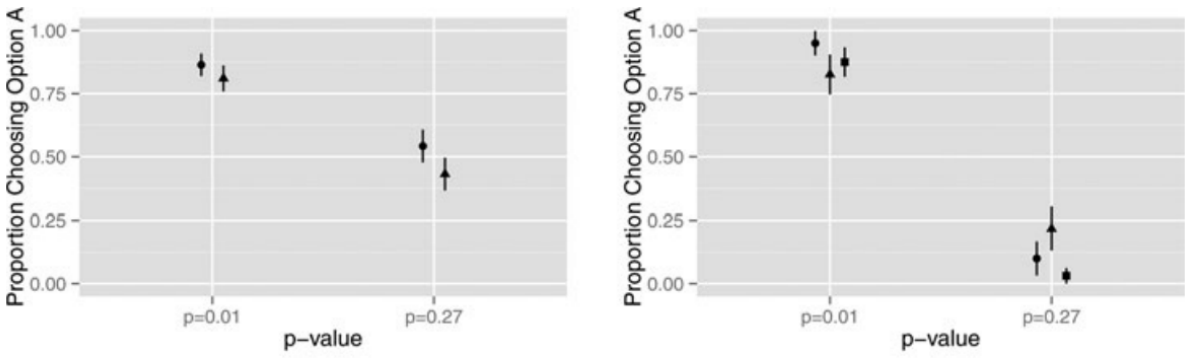
The result was that only at p = 0.025 (statistically significant) did the respondents correctly choose answer "A" a majority of the time in both groups. The treatment difference size was also manipulated, with the small effect being 52% vs. 44% recovery, and the large effect being 57% vs. 39% recovery - and when the effect was larger, the researchers were somewhat more likely to choose "A" in all conditions, as you can see.

Interestingly, in terms of "choice", when the authors said what drug they would choose for themselves, they were much better at choosing "A." But even in the large treatment difference group, it was only around 50-50 for the not statistically significant results.

McShane and Gal's work is interesting because it's an example of academics being wrong in a way that laymen would not be. And this effect is systematic and is apparently caused by their statistical training.

3.5.2 McShane 2017

In 2017[15], McShane and Gal asked the same question we presented to you at the beginning, looking at cancer life expectancies for those who were told to write about their blessings or write about the misfortunes of others. This was asked of 117 authors of articles published in the *Journal of the American Statistical Association*:

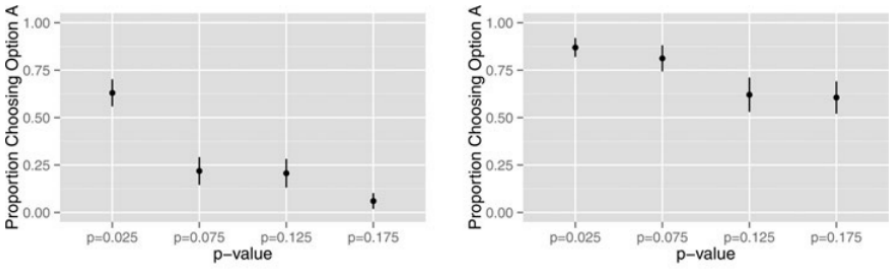


(a) *JASA* (b) *NEJM*

McShane and Gal contrast the results from the *Journal of the American Statistical Association* (JASA) with the results from the *New England Journal of Medicine* (NEJM), and found that, when the results were not statistically significant, the JASA respondents were more likely to correctly answer "A."

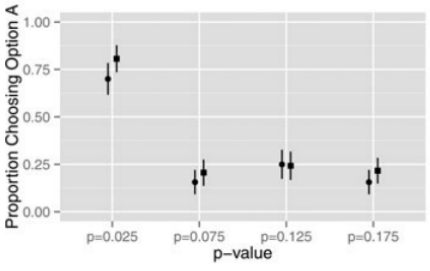
McShane and Gal then repeated the second question from 2016 about the effectiveness of a drug, and compared the results from JASA to the *American Journal of Epidemiology* (AJE):

[15]https://sci-hub.tw/10.1080/01621459.2017.1289846

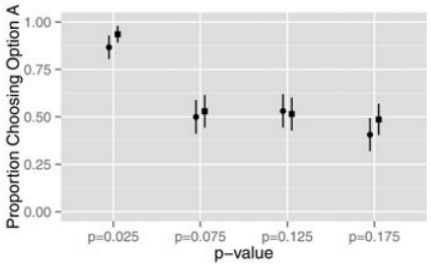


(a) *JASA*: Judgment

(b) *JASA*: Choice



(c) *AJE*: Judgment



(d) *AJE*: Choice

In terms of the proportion choosing option A (Drug A is more likely to have a beneficial effect), the proportion of the JASA respondents who correctly chose answer "A" was only 63% even when the p-value was 0.025, a worse performance than the AJE respondents. At lower p-values, the JASA responses were even more dismal (22%, 21% and 6% respectively).

What these studies show is that epidemiologists and statisticians don't really know what p-values and statistical significance mean in a practical sense, and that there is no important difference between statisticians and epidemiologists on this matter.

3.5.3 Lyu 2019

In 2019 the paper "Beyond psychology: prevalence of p value and confidence interval misinterpretation across different fields"[16], the authors Lyu, Xu, Zhao, Zuo and Hu gave a series of false statements about p-values and confidence intervals to 1,231 mainland Chinese academics, and 248 academics who are Chinese nationals abroad.

These were the results as reported by Lyu et. al:

[16]https://www.cambridge.org/core/services/aop-cambridge-core/content/view/D1520CFBFEB2C282E93484057D84B6C6/S1834490

Table 1. Percentage of misinterpretation of *p* values and CIs for each statement

	Statement (significant scenario)	Science $N = 133$ (9%)	Eng/Agr. $N = 72$ (5%)	Medicine $N = 69$ (5%)	Economics $N = 93$ (6%)	Management $N = 51$ (3%)	Psychology $N = 125$ (8%)	Social Science $N = 111$ (8%)	Math/Statistics $N = 105$ (7%)	Average $N = 759$ (51%)
p value (significant)	(a) You have absolutely disproved the null hypothesis.	53%	53%	49%	60%	63%	50%	59%	44%	53%
	(b) You have found the probability of the null hypothesis being true.	58%	62%	52%	44%	55%	59%	45%	32%	51%
	(c) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.	53%	62%	51%	67%	71%	77%	67%	70%	65%
	(d) You have a reliable experimental finding in the sense that if, hypothetically, the experiment was repeated a great number of times, you would obtain a significant result on 99% of occasions.	62%	54%	64%	63%	53%	42%	59%	48%	55%
	Total (endorsed at least one statement)	93%	90%	90%	92%	94%	95%	95%	88%	92%
CI (significant)	(a) There is a 95% probability that the true mean lies between .1 and .4.	56%	53%	52%	60%	63%	66%	67%	33%	56%
	(b) If we were to repeat the experiment over and over, then 95% of the time the true mean falls between .1 to .4.	59%	56%	54%	54%	51%	54%	59%	48%	55%
	(c) If the null hypothesis is that there is no difference between the mean of experimental group and control group, the experiment has disproved the null hypothesis.	57%	53%	49%	53%	59%	31%	48%	40%	48%
	(d) The null hypothesis is that there is no difference between the mean of experimental group and control group. If you decide to reject the null hypothesis, the probability that you are making the wrong decision is 5%.	62%	53%	48%	66%	63%	70%	56%	58%	60%
	Total (endorsed at least one statement)	97%	93%	93%	96%	98%	94%	94%	88%	94%

Table 1. (*Continued*)

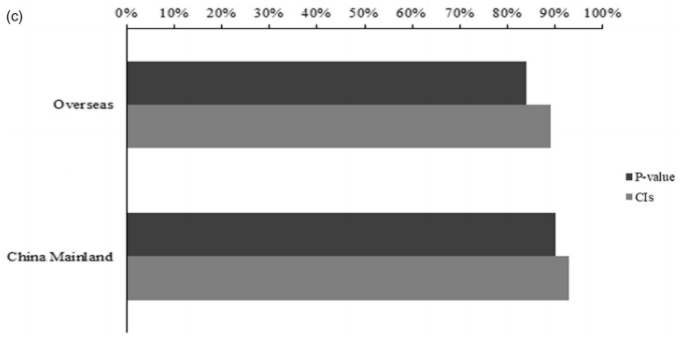
	Statement (nonsignificant scenario)	Science $N = 114$ (8%)	Eng/Agr. $N = 79$ (5%)	Medicine $N = 61$ (4%)	Economics $N = 71$ (5%)	Management $N = 44$ (3%)	Psychology $N = 147$ (10%)	Social Science $N = 106$ (7%)	Math/Statistics $N = 98$ (7%)	Average $N = 720$ (49%)
p value (non-significant)	(a) You have absolutely proved the null hypothesis.	63%	57%	48%	48%	55%	54%	53%	43%	53%
	(b) You have found the probability of the alternative hypothesis being true.	57%	43%	54%	42%	48%	40%	49%	34%	45%
	(c) You know, if you decide not to reject the null hypothesis, the probability that you are making the wrong decision.	54%	56%	64%	65%	70%	63%	59%	55%	60%
	(d) You have an unreliable experimental finding in the sense that if, hypothetically, the experiment was repeated a great number of times, you would obtain a significant result on 21% of occasions.	61%	48%	43%	42%	43%	29%	45%	32%	42%
	Total (endorsed at least one statement)	87%	9%	82%	90%	93%	84%	87%	78%	86%
CI (non-significant)	(a) There is a 95% probability that the true mean lies between -.1 and .4.	62%	54%	62%	61%	55%	69%	63%	33%	58%
	(b) If we were to repeat the experiment over and over, then 95% of the time the true mean falls between -.1 to .4.	53%	49%	52%	56%	61%	48%	60%	53%	53%
	(c) If the null hypothesis is that there is no difference between the mean of experimental group and control group, the experiment has proved the null hypothesis.	54%	44%	61%	46%	43%	46%	50%	37%	48%
	(d) The null hypothesis is that there is no difference between the mean of experimental group and control group. If you decide not to reject the null hypothesis, the probability that you are making the wrong decision is 5%.	52%	58%	51%	51%	68%	53%	63%	45%	54%
	Total (endorsed at least one statement)	95%	92%	92%	89%	98%	89%	93%	85%	91%

I then converted these into average number of errors made by each field for the p-significant, CI-significant, p-insignificant, and CI-insignificant questions:

Question Category	Science	Engineering/ Agronomy	Medicine	Economics	Management	Psychology	Social Science	Mathematics /Statistics	Average
P values (Significant)	2.26	2.31	2.16	2.34	2.42	2.28	2.30	1.94	2.24
CI (Significant)	2.34	2.15	2.05	2.33	2.36	2.21	2.30	1.79	2.19
P values (Insignificant)	2.35	2.04	2.09	1.97	2.16	1.86	2.06	1.64	2.00
CI (Insignificant)	2.21	2.05	2.26	2.14	2.27	2.16	2.36	1.68	2.13
Total	9.16	8.55	8.56	8.78	9.21	8.51	9.02	7.05	8.56

The total number of endorsements of incorrect statements are out of 16 maximum. Within each question category there are 4 false statements. Psychology scored very near the average, scoring 8.51 vs. 8.56 for the whole field, beat out only by the results from Math and Statistics which scored 7.05 out of 16. The similar performance of psychology in comparison to other fields is important, because it makes it implausible that the problems academics have in interpreting statistics is limited to psychology. In fact, by focusing on psychology.

Lyu also showed the total percentage of respondents who endorsed at least one false statement by whether they were Chinese nationals abroad, or if they were Chinese on the Chinese mainland:



The Chinese on the Chinese mainland may be marginally worse than Chinese abroad. But it's also possible that non-Chinese academics in the west are worse at interpreting statistics than Chinese in China. But there's no good reason to reject data on Chinese academics as not being applicable to the west unless there is some very compelling reason to do so.

3.5.4 Zuckerman 1993

In 1993 in the paper "Contemporary Issues in the Analysis of Data: A Survey of 551 Psychologists"[17], Zuckerman et. al looked at the scores of 508 Psychologists, broken down by being Full Professor, Associate Professor, Assistant Professor or Student, on 5 first-year statistics questions. The average number of correct responses by question were as follows:

Table 2. *Mean accuracy scores by question and academic rank*

		Question					
Academic rank	*n*	1	2	3	4	5	Mean
Student	17	.47	.41	.41	.59	.88	.55
Assistant professor	175	.30	.63	.50	.34	.89	.53
Associate professor	134	.35	.63	.66	.43	.94	.60
Full professor	182	.43	.67	.67	.51	.89	.63
Mean (unweighted)	—	.39	.59	.56	.47	.90	.58
Mean (weighted)	—	.36	.63	.60	.43	.90	.59

These questions were all true or false, so random guessing would give a score of 0.5.

3.5.5 Hoekstra 2014

The paper "Robust misinterpretation of confidence intervals"[18] took 594 first year psychology students, master students, and researchers from the University of Amsterdam and gave them six statements about confidence intervals. All six of these statements were false. They then asked the respondents to either endorse or reject these statements.

Table 1 Percentages of students and teachers endorsing an item

Statement	First Years ($n = 442$)	Master Students ($n = 34$)	Researchers ($n = 118$)
The probability that the true mean is greater than 0 is at least 95 %	51 %	32 %	38 %
The probability that the true mean equals 0 is smaller than 5 %	55 %	44 %	47 %
The "null hypothesis" that the true mean equals 0 is likely to be incorrect	73 %	68 %	86 %
There is a 95 % probability that the true mean lies between 0.1 and 0.4	58 %	50 %	59 %
We can be 95 % confident that the true mean lies between 0.1 and 0.4	49 %	50 %	55 %
If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4	66 %	79 %	58 %

[17] https://sci-hub.tw/10.1111/j.1467-9280.1993.tb00556.x
[18] https://sci-hub.tw/10.3758/s13423-013-0572-3

The average number of false statements endorsed by education levels were 3.43 for researchers, 3.23 for master students, and 3.52 for first-year students:

Respondent Type	Number of false statements about confidence intervals endorsed
Researchers	3.43
Master Students	3.23
First Year Students	3.52

In this sample at least, the population that endorsed the fewest false statements were master students, then researchers, then first-year students. On the predictions of effort made by behavioral economists, PhD students performed better than PhDs.

3.5.6 Haller 2002

In a 2002 paper "Misinterpretations of Significance:A Problem Students Share with Their Teachers?"[19], the researchers Haller and Krauss looked at 113 psychology students and professors from 6 German universities, and how they responded to 6 false statements about statistical interpretation - either true or false. They also compared these to the results of a 1986 paper on 70 US "Academic Psychologists" which asked the same questions. The results were as follows:

Group	Average Number of False Statements Endorsed	Note
Methodology Instructors	1.80	
Scientific Psychologists	2.03	
Psychology Students	2.54	
"Academic Psychologists"	2.55	Oakes 1986 - US Psychologists

US Academic psychologists endorsed 2.55 false statements on average in 1986, German psychol-

[19] https://www.researchgate.net/publication/27262211_Misinterpretations_ of_Significance_A_Problem_Students_Share_with_Their_Teachers

ogy students in 2002 endordsed 2.54, "Scientific Psychologists" endorsed 2.03, and Methodology instructors endorsed 1.80 out of 6. Random guessing would result in endorsing 3 false statements on average.

3.6 Nonsense Math

The paper "The Nonsense Math Effect" asked 200 people with post-graduate degrees and asked them what they thought about two articles, one in evolutionary anthropology and one in sociology. They were to give a rating on a scale of 1 to 100. The added nonsense math was this formula here:

> A mathematical model ($T_{PP} = T_0 - fT_0d_f^2 - fT_Pd_f$) is developed to describe sequential effects.

The results were then replicated in a replication study.[20] Both results are shown here:

Table 1

Descriptive Statistics comparison and t-test results

Study		Reanalysis of Original			Replication	
Area of degree	*N*	*M*	*SD*	*N*	*M*	*SD*
Math, Science, Technology	69	-1.28	(19.24	68	3.47	(22.92)
Medicine	16	3.06	(15.99)	18	8.44	(20.57)
Humanities, Social Science	84	**6.60****	(21.15)	136	3.56	(28.86)
Others, e.g. education	31	**13.90****	(23.31)	42	6.86	(25.58)
Total	200	**4.74****	(21.01)	264	**4.39****	(26.33)

Notes. ** p < .01, * p < .05.

The unweighted average rating of the two studies for people whose area of degree was Math, Technology and Science 2.19 percent better than without this formula, for medicine 5.75% better, for the humanities 5.08% better, and education and other fields 10.38% better.

[20] https://osf.io/4t9p2/download

And this is a very small prime. Just a single formula. Consider that the correct response to the additional nonsense meth would be to rate the paper *worse* if the reviewers actually understood everything said in the paper. Instead they rate it slightly better.

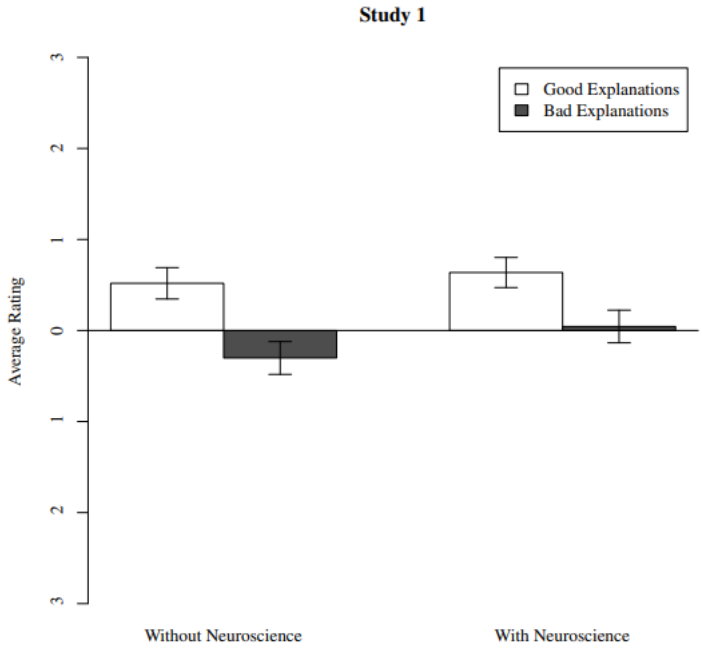
We don't know of any studies looking at the effects of having multiple formulas throughout the paper.

A similar kind of experiment was done in the paper "Deconstructing the Seductive Allure of Neuroscience Explanations"[21].

They had 3 studies for this paper, and were given descriptions of 18 different psychological phenomena. For example, babies' ability to do simple arithmetic, attentional blinking, gender differences in spacial reasoning, differences between seeing and imagining objects.

The subjects were asked to rate the quality of the explanation, and the experimenters made "good explanations" and "bad explanations," but also either added nonsense neuroscience or didn't.

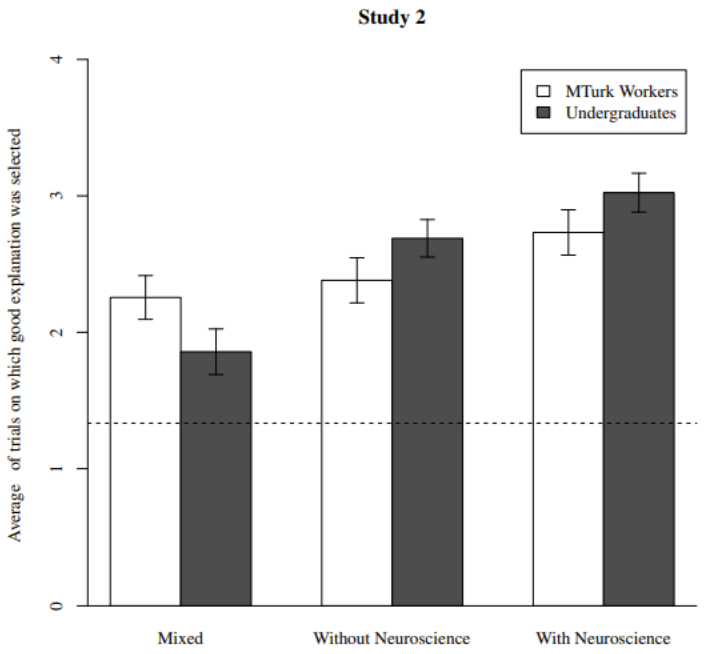
3.6.1 Study 1



From Study 1, we can see the ratings for all participants of good and bad explanations with nonsense neuroscience and without nonsense neuroscience. Good explanations without still beat bad explanations with nonsense neuroscience, but there was still a substantial effect of nonsense neuroscience.

[21] https://repository.upenn.edu/cgi/viewcontent.cgi?article=1143&context=neuroethics_pubs

3.6.2 Study 2

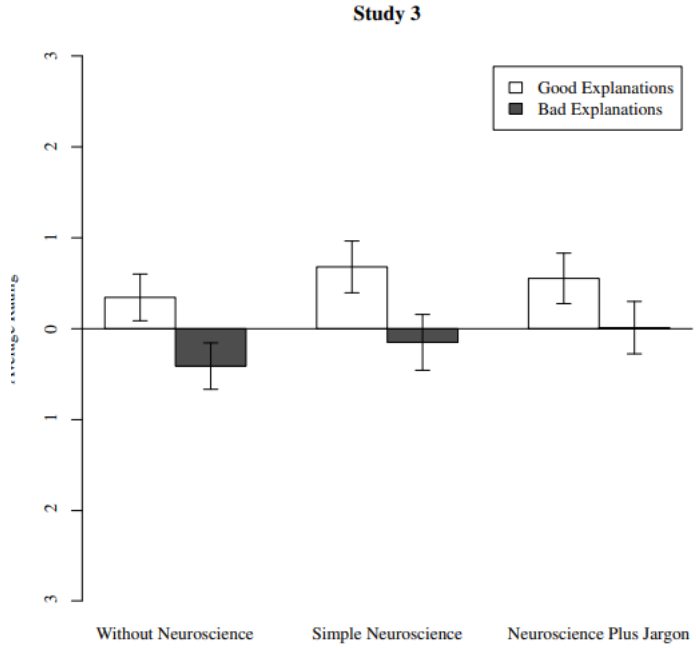


For study 2 the subjects were asked to select the "good explanation" vs. the "bad explanation" when presented both. The without neuroscience condition had good and bad explanations given without any nonsense neuroscience, the with neuroscience explanation had it, and the mixed condition gave them either a bad explanation with nonsense neuroscience or a good explanation without neuroscience.

Those results were interesting. Compared to mechanical turk workers, undergraduates were more likely to select the bad explanation that had neuroscience in it, i.e. to pick the incorrect explanation. Overall however, undergraduates were more likely to pick the good explanation as the good explanation than mechanical turk workers.

But in a situation where the wrong explanation has more nonsense science to seemingly back it up, randos did better than undergrads.

3.6.3 Study 3



In study 3, the researchers against asked for ratings of explanations of phenomena giving good and bad explanations for all 3 groups. One was without neuroscience, one was with simple neuroscience, and one was with neuroscience plus jargon.

The without neuroscience explanations were rated the worst, the simple neuroscience was better for both good explanations and bad explanations, but adding in jargon helped the rating of a bad explanation the most. It appears adding jargon has a compression effect, reducing the perceived quality of a good explanation but elevating the perception of an otherwise bad explanation.

To the extent some academics are aware of this phenomenon, either explicitly or implicitly, they may resort of pumping up the jargon if their case is weak.

4 The Journal System

4.1 Word Game

The First problem with Peer Review is the term "peer review." In reality, all scientific papers that are written by anyone are in fact peer reviewed - in the sense that they are reviewed by their peers and colleagues. Whether intentional of not, the very label "Peer Review" conveys an inaccurate idea about what is being discussed. Nobody is against peer review; the debate surrounds the efficacy and honesty of *"Peer Review"* - with a capital P.

In addition, the word "peer" in Peer Review is redundant, because who would review the paper other than one's peers if we're operating at the cutting edge of science? If a researcher is among

the top researchers in their field, and there is nobody generally considered to be above them in their overall knowledge, then the only possible reviewers could be their peers.

"Peer" is redundant, and everything is "reviewed" whether it goes through a formal journal system or not. And so the issue at hand is the journal system and the review boards of the journal systems specifically. Not some argument about reviewing vs. not-reviewing work as if one side is against review.

4.2 Basic Knowledge Problem

If you spend say, 6 months studying a very specific topic, say insulin regulation in muskrats, there is a very good chance that you know more about insulin regulation in muskrats than anyone in the world.

Who then, should review your paper? Maybe there's another guy from Japan who's studied insulin regulation in groundhogs for a similar period of time. In this case, it makes perfect sense to have your paper reviewed by *that guy*. Maybe he knows some things you don't, maybe he can either find an error or see that you're trying to solve a problem in an inefficient way.

That makes perfect sense. And your "review" would involve exchanging notes, long Skype calls with the guy from Japan who studied insulin regulation in groundhogs.

Where does *Nature* or *The Lancet* come into play? None of the people at these big journals know anywhere near as much about insulin regulation in muskrats or other mammals as you or that guy from Japan. But it has become a convention that papers *must* be published in these big journals. And so you send your paper to Nature, and they set up a review board. Unless the review board just so happens to include that guy from Japan, none of the people reviewing your paper has anywhere near the knowledge of insulin regulation in muskrats that you do. So what's the point?

This is the basic knowledge problem of the journal system.

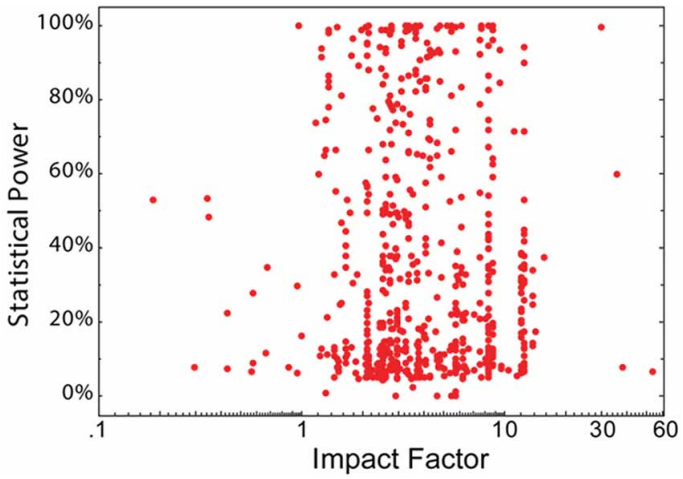
4.3 Big vs. Small Journals

The paper, "Prestigious Science Journals Struggle to Reach Even Average Reliability"[22], Bjorn Brembs looks at various aggregate proxies for article quality that could be applied over a large number of articles - and see if higher impact journals have "higher quality" papers. Of course, how "good" a paper is is subjective, but you can do things like look at the statistical power of hundreds of papers, and if one journal repeatedly has lower statistical power that is at least a *sign* that a journal has "lower quality" papers generally.

It is *some* objective hook we can latch onto in what is mostly a subjective process.

And from this, we can look at the average statistical power in neuroscience and psychology papers by the impact factor of the journal they are published in.

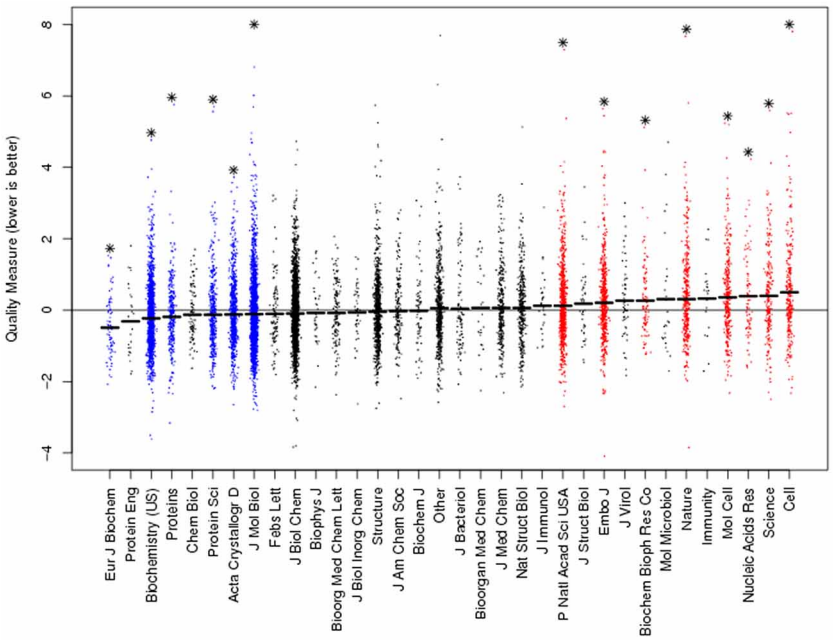
[22]https://www.frontiersin.org/articles/10.3389/fnhum.2018.00037/full



Impact factor being a measure of the journal's "prestige" calculated by citations of articles from that journal. The more cited, the higher the impact factor, the more "prestigious" the journal is. (This is of course another problem in that the journal system is, definitionally, a self-referential prestige system.)

Brembs looked at 730 studies, and found no relation between journal rank and statistical power of the study.

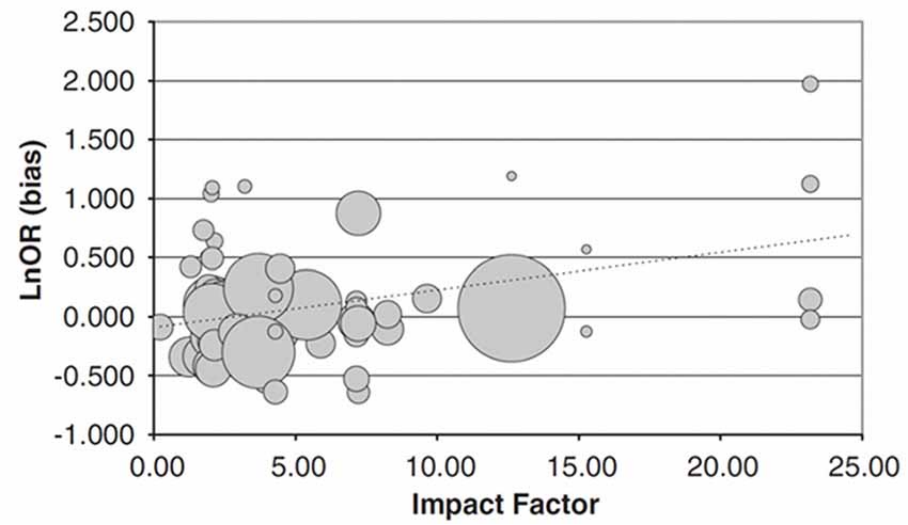
Another method Brembs used, which is another "objective hook," was to look at crystallographic quality - or the quality of models used in crystallographic work, and seeing how often they deviate from known atomic distances. Brembs looked at an analysis by Brown and Ramaswamy, which looked at 17,503 structures from papers in 30 journals. Brembs then just looked at the quality ratings of each journal by impact factor at the time of Brown and Ramaswamy's analysis:



And what Brembs found was that higher impact factor journals had, on average, worse crystallographic work than lower impact-factor journals.

You can say this is a rather limited indicator of a journal's quality, but it is at least *some* indicator.

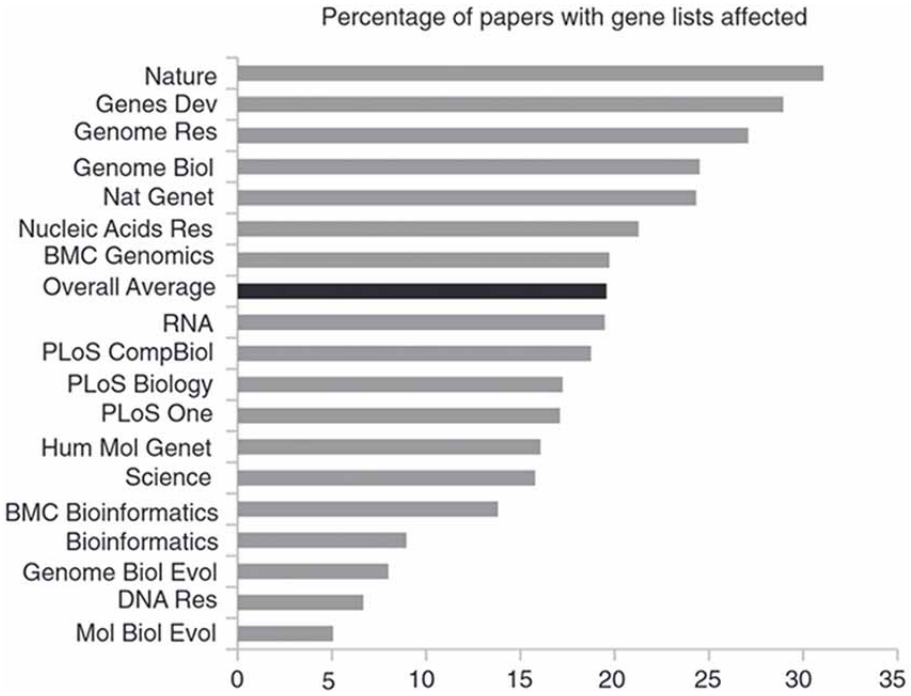
Brembs also looked at gene-association studies, citing a meta-analysis of meta-analyses done by Munafo. Munafo looked at the effect size of a gene analyzed in individual studies, and compared them to how much they deviate from meta-analyses of that gene.



The higher the number here [LnOr (bias)], the more the individual study published deviates from the result of ensuing meta-analyses of the effect size of that gene. The large the circle, the larger the population sample size of that study.

What Brembs found based on Munafo's work was that high impact factor journals had studies with smaller sample sizes, large effect sizes, which were found to deviate from the results of later meta-analyses. At least in the realm of gene-association studies. Deviation from the results of a meta-analysis is treated as evidence that an individual study is wrong, under the assumption that the pooled estimate is closer to the truth. And so this is more of a soft refutation - it could be that the study with the small sample size and bigger effect is closer to the truth and everyone else is wrong, especially if there is some methodological dispute, but given the problem of induction the convention is to generally take the results of meta-analyses over the results of a single study that deviates from the meta-analysis.

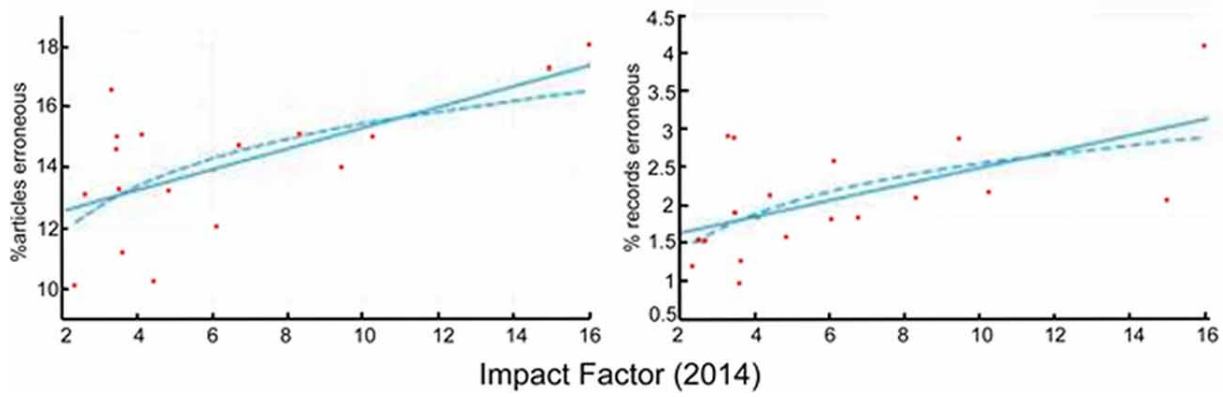
Brembs also showed the work of Ziemann 2016, which looked at how often papers got gene symbols incorrect in their papers:



This is a very simple and straightforward objective hook. The gene symbol is just the name, for example BRCA1. And Ziemann looked at 3,597 papers and found that In about 20% of all papers, a gene symbol was used incorrectly somewhere. Again with high impact factor journals having the most errors - *Nature* itself having the worst.

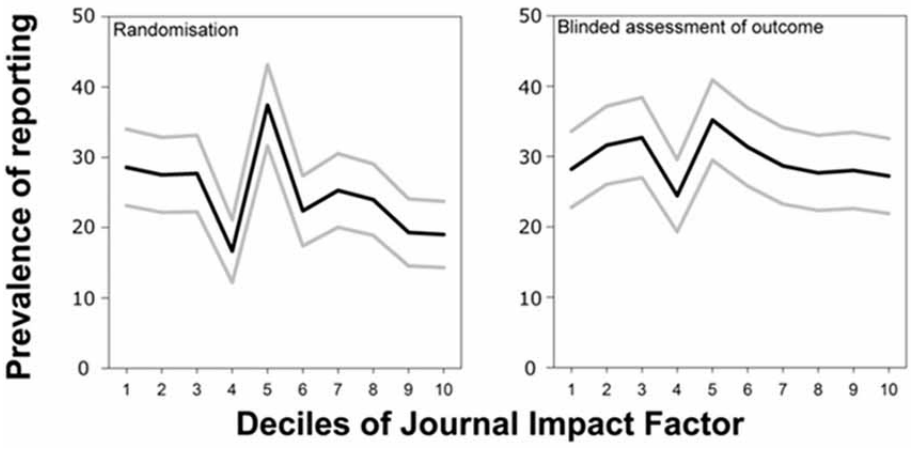
To repeat, it's not that mislabeling genetic data here or there is a huge deal, but it's another objective hook which shows that the more prestigious journals - if anything - have more of this basic error than the less prestigious journals do.

Brembs analyzed the work of Szucs and Ioannidis on the rate of miscalculated p-values, and organized their data by journal impact factor. Brembs found that the higher impact factor journals were more likely to publish an article with a miscalculated p-value:



Brembs also cited the work of Macleod, which looked at 814 randomly selected English-language

papers involved in primary research. In it, he looked at how many papers engaged in blinding and how many used a control group.



Macleod found that higher-impact journals had roughly the same amount of blinding as low-impact journals, but a lower amount of randomisation.

Fittingly, Brembs himself mis-cited the Macleod paper, citing a correction of the spelling of one of the author's names instead of the paper itself. A fitting mistake for Brembs' analysis using technical errors as an objective hook for a paper's quality. And of course, along with all of these other papers, Brembs' paper passed "peer review" with this little error intact.

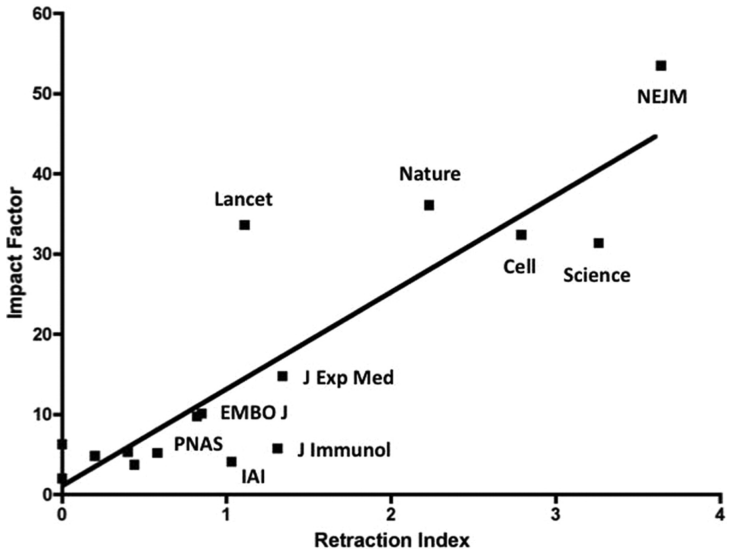
Now look, all of this may seem petty if you don't understand what the point is. The point here is what was stated at the beginning - there's no reason to think that lower impact factor journals produce lower quality papers than higher impact factor journals, as a general rule. They could, but you have no good reason to think that.

In the absence of *any* evidence, you may think Nature and Lancet publish higher quality papers because they're more prestigious. And nobody should fault you for thinking that *if you had no other evidence to go on*. But now you do. You could go on thinking lower impact factor journals are in fact more rigorous, and think up some reasons for that, reasons which may or may not be true, and in fact the lower impact factor journals may or may not actually be more rigorous.

The point being made here is more limited than that.

The point of this section is to cement something that you may not think is important now - which is that there is no reason to think that high impact-factor journals are any more diligent than low impact-factor journals. And it's perfectly fine if you think this is a small or unimportant point right now. But when we start looking at journal stings - prestige stings, error detection stings, results stings, SciGen paper acceptance - these use journals of opportunity. Rarely do the big name journals agree to such stings.

Last point on big vs. small journals: larger journals have a higher proportion of retracted papers. From Fang and Casadevall, they looked at journals, the percentages of papers were retracted, and came up with a retraction index:

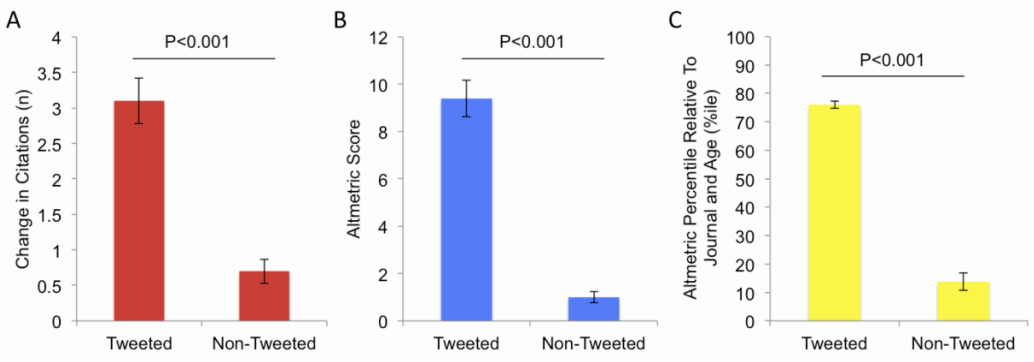


We shouldn't read too much into this particular measure, as the higher impact factor journals may just be getting more scrutiny. However, in order to say "the bigger journals are better," you have to have some sort of explanation for their higher rate of retractions.

4.4 Article Prominence

From the paper "Does Tweeting Improve Citations? One-Year Results from the TSSMN Prospective Randomized Trial"[23], the researchers took 112 articles from 2017-2018 from *The Annals of Thoracic Surgery* and the *Journal of Thoracic and Cardiovascular Surgery*. 56 of them were left alone, 56 were retweeted by a twitter account with 52,893 followers at the time of the study.

The randomisation was weighted such that the articles selected to be tweeted and those not selected to be tweeted had the same number of citations before the experiment, the same average age, and were distributed among the same topics within the journal.



And the results were that the tweeted articles had a roughly 6-fold increase in citations, 10-fold

[23]https://sci-hub.tw/10.1016/j.athoracsur.2020.04.065

increase in Altmetric score, 5 fold increase in Altmetric percentile controlled for journal impact factor and age of the article.

This study is just looking at the effect of twitter, but in the absence of any other data on other media reports, there's no reason to assume the effects from any other kind of media exposure that increases the prominence of an article among the general public wouldn't also have similar effects on citation rates among research scientists.

And while you may have never heard of *The Annals of Thoracic Surgery* and the *Journal of Thoracic and Cardiovascular Surgery* - and perhaps it's a small journal - there's no reason to imagine it's any worse than the big journals.

The point being that what researchers know about is influenced by media reports. And in fact this effect can be much larger than anything to do with the content of research papers themselves.

4.5 Outcome Stings

From the paper "Reviewer Bias. Annals of Internal Medicine"[24], from Ernst, Resch and Uher, the authors sent out a fictitious paper on the effectiveness of electrical nerve stimulation to 33 reviewers. The fictitious papers were all identical with the exception of the results: they were either positive or negative. The reviewers were then asked to rate the papers on 5 factors - study design, patient descriptions, statistical methods, end points and linguistic quality on a scale of 1-5.

"Effect of Blinded Peer Review on Abstract Acceptance"

	Open	Double-Blind	Double-Blind "True Estimate"
High Prestige	51.3	38.8	30.12
Moderate Prestige	42.7	34.3	28.46
Low Prestige	32.6	29.0	26.49
High Prestige/ Low Prestige	1.574	1.338	1.137

Based on the raw scores, the "positive results" papers had an average score 39.54% greater than the negative results. So the otherwise identical paper was rated much higher if it had positive results. This is evidence of the "positive results bias."

— Warning - big digression here —

However, there is a restriction of range, meaning the lowest score on each factor is 1, and so the lowest possible score is 5, the highest possible score is 25.

[24] https://sci-hub.tw/10.7326/0003-4819-116-11-958_2

Imagine if an author decided to ask reviewers to review their book, and had them rate the book on a scale of 9 to 10, with 9 being the lowest and 10 being the highest. And then advertised how the reviewers all gave his book 9/10 reviews - a ringing endorsement! And then if a competing author had his books reviewed on the same 9 to 10 scale, and got all 10/10 reviews. Well, would we say the second author's books only got 10% better reviews? This is the problem of range restriction, and why you must analyze the variation within the available range of scores.

That is the problem here to a lesser extent. The lowest score needs to be zero. So if this was done on a scale of 0 to 4 with a maximum score of 20, the negative results papers would have a score of 7.39, the positive results paper would have an overall score of 12.29, or a 66.3% higher scores.

—-

But the takeaway is that, in this experiment at least, the *results* mattered in terms of how reviewers evaluated the methodology. If they disagree with the result, they are more likely to say the methodology is poor, independent of how good the methodology actually is.

A similar manuscript sting was done by Epstein in 1990. He submitted 146 papers dealing with social work. 86 received a response by the time Epstein published his paper.

Table 3. Publication Decisions Among "Relevant" Social Work Journals

	Positive version		*Negative version*	
Decision	*N*	*%*	*N*	*%*
Accept for publication	6	35.3	4	25.0
As is or minor revisions	5		2	
Moderate or extensive	1		2	
Possible acceptance	2	11.8	0	0.0
Reject for publication	9	52.9	12	75.0
Not Relevant	4		2	
Substantive reasons	3		6	
Both	1		2	
No reasons provided	1		2	
Total Reviewed	17	100.0	16	100.0
Decline to review	10		12	
Irrelevant	4		8	
Other reasons	6		4	
Total	27		28	

TABLE 4. Responses of "Allied" Journals

	Positive version	*Negative version*
Response	*N*	*N*
Total sample	16	15
No response	1	0
Declined to review	8	10
Accepted for review	7	5
Accepted for publication	2	0
Possible acceptance	1	2
Rejected	4	3

The combined results for the social work and what Epstein classified as "allied" journals were as follows:

For the negative result, 21 were accepted for review. For the positive result, 24 were accepted for review. Minor variance that could just be randomness.

For the negative result, 4 were accepted for publication, 8 for the positive.

For possible acceptance, 2 for the negative, 3 for the positive.

22 of the negative result papers were rejected outright, 17 of the positive results rejected outright.

Keep in mind thresholds. The Ernst paper looked at how reviewers rated the paper. A reviewer may give one paper a lower score than another - but still decide to publish both. For example, one paper may be rated a 9, the other paper rated an 11, and the reviewer may recommend publication for both. If you just look at decision to publish or review, the variation in scores *within* those thresholds is ignored, meaning less bias is captured.

You can think of it how guys rate girls, and whether they would have sex with them. They may rate one girl a 7 and another girl a 9, but would have sex with either of them. And so the decision to have sex by definition doesn't give any information above the threshold of girls a guy is willing to have sex with. Same with the thresholds of deciding to review or publish.

And thus the results of the Epstein paper are a more profound indicator of bias than would initially appear - because it implies that the negative results papers get rated so much more poorly that they fall below the thresholds of acceptance for review and publication outright at a higher rate.

From the paper "Testing for the Presence of Positive-Outcome Bias in Peer Review"[25], the authors took a paper purportedly on a randomized controlled trial on the efficacy of a form of knee joint surgery. Again, two versions of the paper - one with positive results and one with negative results. The paper was identical in every aspect except the results, and the paper was sent to 238 reviewers.

Table. Rates of Reviewers' Recommendations for Acceptance, Error Detection, and Methods Scores of Manuscripts With Positive vs No-Difference Findings at 2 Orthopedic Journals

Journal	Positive Version, No./Total No. (%)	No-Difference Version, No./Total No. (%)	*P* Value	OR (95% CI)
		Accept Manuscript		
CORR	58/60 (96.7)	43/48 (89.6)	.28	3.37 (0.62-18.21)
JBJS	49/50 (98.0)	37/52 (71.2)	.001	19.87 (2.51-157.24)
Total	**107/110** (97.3)	**80/100** (80.0)	<.001	8.92 (2.56-31.05)

	Error Detection				
	Positive Version		**No-Difference Version**		
	Reviewers, No.	**Score, Mean (SD) [95% CI]**	**Reviewers, No.**	**Score, Mean (SD) [95% CI]**	***P* Value**
CORR	60	0.52 (0.68) [0.29-0.75]	48	1.00 (1.34) [0.74-1.26]	.02
JBJS	50	0.28 (0.45) [0.03-0.53]	52	0.71 (0.96) [0.47-0.96]	.005
Total	**110**	0.41 (0.60) [0.23-0.57]	**100**	0.85 (1.16) [0.68-1.03]	<.001
		Methods Scores			
CORR	60	7.87 (1.81) [7.38-8.36]	48	7.38 (2.37) [6.83-7.93]	.22
JBJS	50	8.68 (1.21) [8.14-9.22]	52	7.66 (2.17) [7.14-8.19]	.005
Total	**110**	8.24 (1.61) [7.91-8.64]	**100**	7.53 (2.26) [7.14-7.90]	.005

Abbreviations: CI, confidence interval; *CORR, Clinical Orthopaedics and Related Research*; *JBJS, The Journal of Bone and Joint Surgery*; OR, odds ratio.

[25]https://sci-hub.tw/10.1001/archinternmed.2010.406

For the positive results paper, the acceptance rate was 97.3%, for the negative results paper the acceptance rate was 80%.

There were 7 intentional errors planted in the paper. For the positive result papers the reviewers found on average 0.41 of the intentionally planted errors, the reviewers of the negative results paper found on average 0.85 errors. This is an error detection rate of 5.86% and 12.14% respectively.

For the method scores, the positive results papers scored 8.24, the negative results papers scored 7.53.

Now the method score gap may seem small, but keep in mind the variance indicated by the standard deviations. Based on the standard deviation of the positive results method scores, the negative results method scores were are at the 32.96th percentile. If there was no bias, they would be at the 50th percentile:

Keep in mind that all of these studies are on how the paper is rated methodologically, and whether to publish. It doesn't measure whether the reviewers changed their minds. That is something these sting operations are incapable of testing - since the papers are fictitious and if a major revision movement within the field started gaining steam, these papers would get greater scrutiny and be found to be fictitious. Thus this method only tests the acceptance rate and initial methodological evaluation at the point of article submission.

It's merely an evaluation of whether the heterodox view gets a fair hearing in the first instance - the answer being mostly no; changing the view of the field itself would be a task unto itself on top of that.

4.6 BMJ Error Detection Sting

In 2008, the British Medical Journal did a sting looking at error detection rate for reviewers at the BMJ. They gave 607 reviewers one of three papers, each which contained 9 intentionally planted "major" errors, and 5 intentionally planted "minor" errors. Any additional errors in the paper that were not intentionally planted (but were genuine errors on the part of the authors) were not used in the analysis so that the evaluation of each papers could be compared more plainly.

Table 4

Mean (SD) errors identified by group for each paper

	Major errors	Minor errors
Paper 1		
Control group (n=173)	2.38 (2.0)	0.99 (0.9)
Self-taught group (n= 166)	2.68 (1.7)	0.79 (0.8)
Face-to-face group (n= 183)	2.68 (1.8)	0.94 (0.8)
All groups combined (n= 522)	2.58 (1.9)	0.91 (0.8)
Paper 2		
Control group (n= 162)	2.13 (1.6)	0.71 (0.8)
Self-taught group (n= 120)	3.14 (1.4)	1.05 (0.9)
Face-to-face group (n= 158)	2.96 (1.7)	0.84 (0.8)
All groups combined (n= 440)	2.71 (1.6)	0.85 (0.8)
Paper 3		
Control group (n= 156)	2.71 (1.8)	0.96 (0.9)
Self-taught group (n= 111)	3.37 (1.7)	1.21 (0.8)
Face-to-face group (n= 151)	3.18 (1.8)	1.12 (0.8)
All groups combined (n= 418)	3.05 (1.8)	1.09 (0.8)

As part of the study, the BMJ either did nothing and had the reviewers just review the paper (which is what happens in the real world), gave them an information packet designed to help them find the most common errors in scientific research, or had a face-to-face session where the reviewers were taught by another person these things.

The unweighted average number of errors detected for the major errors from the three papers were 2.41 for the control group, 3.06 for the self-taught group, and 2.94 for the face-to-face group. This is an error detection rate of 26.8%, 34% and 32.67% respectively.

For the minor errors, the unweighted average of errors detected was 0.89 for the control, 1.02 for self-taught, and 0.97 for face-to-face training. This is an error detection rate of 17.73%, 20.33% and 19.33% respectively.

Now is finding these kinds of errors the hallmark of "good quality science"? Again, peer review is a subjective process. But it's another objective hook into the otherwise subjective process. Which is that they're not particularly good at finding errors.

4.7 Prestige Stings

From the paper "Reviewer bias in single-versus-double-blind peer review. Proceedings from the National Academy of Sciences"[26], the authors looked at the effect of various factors on the ratings of papers submitted to the 10th Association for Computing Machinery International Conference on Web Search and Data Mining, which had an overall acceptance rate of 15.6

[26]https://sci-hub.tw/10.1073/pnas.1707323114

The reviewers were put into two categories - single-blind and double-blind. Single-blind in this context means that the reviewers know who the authors are, but the authors don't know who the reviewers are. Double-blind means that the reviewers also don't know who the authors are.

The researchers compared the scores of the single-blind to the double-blind reviews so they could see how various facts about the authors of the papers effected paper acceptance or review score compared to when the reviewers didn't know who the author was.

They looked at the effect of the "blinded paper quality score" (bpqs), which was how well the paper scored when the author wasn't known. The prestige of the company the author was from if any, how famous the author is as determined by the author's impact factor which is largely a function of citation rate, the rated prestige of the university the author is from, whether the author is a woman, whether the author is from the same country as the reviewer, whether the author is an academic, and whether the author is from the united states.

The same papers were sent to the reviewers who knew who the authors were, and the reviewers who didn't. Blinded vs. unblinded reviewers. Since the unblinded reviewers know how famous an author is, know if he works for a big company, know if he's from a prestigious university, know if he's not a he but a she, we can see the effect of these factors on how the paper is judged compared to the blinded reviews.

These results are presented in Table 2:

Table 2. Learned coefficients and significance for review score prediction

Name	Coefficient	SE	Confidence interval	*P* value	Odds multiplier	bpqs equivalent
Const	−1.83	0.24	[−2.31, −1.36]	0.000	0.16	—
bpqs	0.80	0.08	[0.64, 0.97]	0.000	2.23	1.00
Com	0.74	0.24	[0.27, 1.21]	0.002	2.10	0.92
Fam	0.49	0.22	[0.05, 0.93]	0.027	1.63	0.61
Uni	0.46	0.18	[0.09, 0.83]	0.012	1.58	0.57
Wom	−0.25	0.18	[−0.60, 0.10]	0.160	0.78	−0.31
Same	0.14	0.24	[−0.34, 0.62]	0.564	1.15	0.17
Aca	0.06	0.22	[−0.38, 0.51]	0.775	1.07	0.08
United States	0.01	0.21	[−0.42, 0.44]	0.964	1.01	0.01

The most interesting things to look at here are the coefficient and bpqs equivalent.

The coefficient tells us how much a 1 standard deviation in a factor (say, author famousness) effects review score in standard deviations. For example, if an author is 1 standard deviation more famous than the mean, his paper will be rated 0.49 standard deviations higher than if this information wasn't known to the reviewers.

And if a paper scored 1 standard deviation higher in among the blinded reviewers, that translates to a 0.8 standard deviation higher score among unblinded reviewers.

However, other factors add up to be more important than the score one's paper got in the blinded setting. Whether you work for a prestigious company, hail from a prestigious university, are an author that gets cited a lot, aren't a woman - these factors taken together end up being more important than how your paper would score in a blind review.

The bpqs, or blinded paper quality score equivalent, shows what a 1 standard deviation increase in one of these factors translates to the equivalent effect in standard deviations of a blinded paper quality score.

For example, being 1 standard deviation more famous than the average has if your paper was in fact 0.61 standard deviations "better" according to how the paper was reviewed blinded. Being from a university 1 standard deviation more prestigious than the mean has an equivalent effect on your review score as a 0.57 standard deviation higher review score if you paper were reviewed blind.

So if you work at a big company like Apple or Google, are a famous author, went to MIT, are a man, are from the same country as the reviewer, and are an academic - well your paper would have to be unbelievably horrible to not get a good review.

This analysis is particularly good because it breaks down the factors. Other analyses on blinded vs. unblinded review just look at the overall effect. And this can tell us something about the bias, but it only tells us the average effect.

For example, the paper "Single-blind vs. Double-blind Peer Review in the Setting of Author Prestige"[27] just compares the double-blind vs. single-blind group. The give the same papers to two groups of reviewers, one knows the identity of the authors and one doesn't.

Table 2. Reviewer Scores and Number of Errors Detected for Single-blind vs Double-blind Peer Review

	Mean (SD)			
	Double-blind Group	Single-blind Group	Difference (95% CI)	*P* Value[b]
Reviewer score (range, 0-10)[a]				
Overall score	5.71 (2.18)	7.06 (2.09)	1.35 (0.56 to 2.13)	<.001
Originality of problem	6.70 (2.21)	6.98 (2.09)	0.28 (−0.5 to 1.07)	.49
Methods	6.05 (2.08)	6.97 (2.10)	0.92 (0.15 to 1.68)	.02
Results	6.41 (2.00)	7.23 (2.08)	0.82 (0.07 to 1.57)	.03
Discussion				
Limitations	5.93 (2.12)	6.97 (2.26)	1.04 (0.24 to 1.84)	.01
Literature review	6.40 (1.82)	7.42 (1.76)	1.02 (0.36 to 1.68)	.003
Organization	6.98 (1.75)	7.87 (1.64)	0.89 (0.26 to 1.52)	.006
Clarity of tables and figures	6.45 (1.98)	7.40 (1.53)	0.95 (0.27 to 1.62)	.006
No. of errors detected (maximum of 5)	0.61 (0.77)	0.90 (0.94)	0.29 (−0.02 to 0.60)	.07

And while you can see that when the reviewers know who the author is, they do give higher scores, they don't break down how big the effect can truly be, because it's averaged. The reviewers can know who the author is, one paper could be from North Carolina A&T, the other could be from Cal Tech, and the effect of institutional prestige can cancel out. So what we're seeing here is

[27] https://sci-hub.tw/10.1001/jama.2016.11014

just the net effect of author famousness, institutional prestige, et cetera.

And another problem with this is that, if you had a sufficiently representative sample of papers, you could find no *overall* effect from blinding reviewers to the author's identity, because presumably, the negative effect of low prestige and the positive effect of high prestige roughly cancel out.

And you could use this to say there's no effect of prestige when there's actually a *huge* effect from these prestige factors, they just go positive and negative and if pooled together they'll cancel out.

For this reason we think most single-blind vs. double-blind analyses are flawed and don't give a good idea of how powerful prestige effects *can* be.

That said, even with this flaw, among these papers, knowing who the author was were caused their paper to go from a score of 5.71 out of 10 to a 7.06 out of 10.

And you see their organization was rated higher, the clarity of tables and figures was rated higher, the literature review was rated better, the discussion of study limitations was better. Interestingly though, when they knew who the author was, they detected slightly more errors, 0.9 vs. 0.61 out of 5 on average. Again, pretty abysmal but we've already been over the abysmal error detection rate in journals. Maybe they detected more errors when they knew who the author was because they were more enthralled by the paper, maybe they know more about a particular author's idiosyncrasies, or maybe it's just noise.

4.8 Fake Papers

In 2005, three MIT graduate students Jeremy Stribling, Dan Aguayo and Maxwell Krohn wrote the program SCIgen to generate fake papers. In their sting, they submitted a paper to the 2005 World Multiconference on Systemics, Cybernetics and Informatics. That paper was entitled "Rooter: A Methodology for the Typical Unification of Access Points and Redundancy"[28].

Here's the abstract from the fake paper:

> "Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over-IP and public private key pair. In order to solve this riddle, we confirm that SMPs can be made stochastic, cacheable, and interposable. . . "

The three authors were invited to speak at the conference, where they exposed the hoax. The program SCIgen is available on the internet free to download and use by anyone.

In 2013, at least 16 SCIgen papers have been found in Springer journals.

According to the paper by Dominique and Cyril Labbe entitled "Duplicate and Fake Publications in the Scientific Literature: How many SCIgen papers in Computer Science?"[29], SCIgen papers had an acceptance rate of 13.3% at the ACM digital library, and 28% for Institute of Electrical and Electronics Engineers.

Now certainly the ACM digital library and the IEEE are not the most prestigious journals. But 16 got into Springer. Now we don't know what percentage of SCIgen papers got in, but some did. And if completely bogus and ridiculous nonsense-jargon papers could get in at least some of time, what about papers which aren't so transparently bogus? Whose authors are smarter liars than a text-spinning algorithm?

[28] https://pdos.csail.mit.edu/archive/scigen/rooter.pdf

[29] https://hal.archives-ouvertes.fr/file/index/docid/713555/filename/0-FakeDetectionSci-Perso.pdf

This is the point. Nobody would say that the prestigious journals are literally churning out thousands of SCIgen papers, but the fact that sometimes SCIgen papers can get through calls into question the seriousness of the peer review process.

Another sting operation was done by John Bohannon. Bohannon wrote essentially the same paper 304 times about some moss that inhibited cancer growth. The paper has glaring flaws that he describes in his Sciencemag article, "Who's Afraid of Peer Review"[30].

Among them were descriptions of a correlation between moss exposure and cancer inhibition when his own chart showed zero correlation. He posed as researchers from various third-world institutes, using randomly generated names for the authors and institutions of his 304 fake papers, and moving paragraphs around.

These are the same text "spinning" techniques used by spammers to get past spam filters. He also ran his original text through google translate into French, and then back into English, and then manually corrected the biggest errors in the final translation. This was so he had the correct grammar, but the idiom of a foreign speaker.

The 304 slightly different papers were sent to 304 Journals. In total, 157 were accepted, 98 rejected, 29 were derelict, and 20 were still reviewing the paper by the time Bohannon published the results of his sting.

He sent the paper to 167 Directory of Open Access Journals (DOAJ), and 121 to Jeffrey Beall's list, and 16 on both Beall's list and the DOAJ.

Beall's list is a list of Journals determined by Jeffrey Beall to be bogus. The Directory of Open Access Journals is run by Lars Bjørnshauge, a library scientist at Lund University in Sweden.

Bohannon says of the DOAJ,

> "Without revealing my plan, I asked DOAJ staff members how journals make it onto their list. "The title must first be suggested to us through a form on our website," explained DOAJ's Linnéa Stenson. "If a journal hasn't published enough, we contact the editor or publisher and ask them to come back to us when the title has published more content." Before listing a journal, they review it based on information provided by the publisher."

The results of the sting were as follows:

Reaction	DOAJ	Beall's List	Overlap
Rejected w/o peer review	44.4%	3.1%	3 (total)
Rejected with peer review	11.1%	10.3%	2 (total)
Accepted w/o peer review	24.3%	48.5%	6 (total)
Accepted with peer review	20.1%	38.1%	3 (total)
Total responses	144	97	14

The fact that "junk journals" accepted a junk article is not interesting. Not because these journals are actually worse - they may or may not be - but because of public perception that they are worse.

[30]https://science.sciencemag.org/content/342/6154/60.full

What is interesting is that journals run by Sage, Elsevier and Wolters Kluwer all accepted Bohannon's bogus paper.

Sage's journal named *Journal of International Medical Research* accepted the paper,

Wolters Kluwer's journal *Journal of Natural Pharmaceuticals* accepted the paper, and

Elsevier's journal *Drug Intervention Today* accepted the paper.

Springer, Sage, Wolters Kluwer and Elsevier all went into damage control mode with apologies and statements.

For example, Elsevier says that they don't actually own *Drug Intervention Today*. The problem though is that it's published by Elsevier, and anyone who reads something from *Drug Intervention Today* will see right up top a big "Elsevier" logo on it because it's published right along with Elsevier's other journals. The fact that they don't legally own the journal is a red herring; and this distinction was only highlighted by Elsevier when it got caught in this sting.

Same with Wolter Kulwer's *Journal of Natural Pharmaceuticals*. Wolters Kluwer shut down that journal in response to this sting. But there's no reason to believe that the *Journal of Natural Pharmaceuticals* was any worse than any of Wolter Kluwer's other journals. That just happened to be the journal targeted by Bohannon's sting.

Bohannon's sting and the SCIgen sting show that horrifically bad papers can get through with some regularity.

5 The Great Stagnation

5.1 Economic "stagnation"

If you do a Google or DuckDuckGo search for "the great stagnation," you will often find it paired with the idea that "the low-hanging fruit has been plucked." This is certainly possible, and even if you don't think it's the primary cause of some great stagnation, it may still be a factor.

So the first thing we should look at, because it's the easiest to measure, is the change in real wages. Measurements of real wages before 1979 are spotty, but there are some estimates that go back further. But "real wage" is dependent on a price index. The most well-known price index is the Consumer Price Index (CPI), which tracks the change in nominal price in a basket of goods, and compares that to the change in nominal wages to get "real wages."

Another metric that is gaining popularity is the Personal Consumption Expenditure (PCE) index. This looks at how much individuals are spending, and instead of using household surveys on the prices of things, they look at business surveys on the sticker price of things. Now survey does not mean "poll," but actual receipts. The CPI and PCE surveys the receipts.

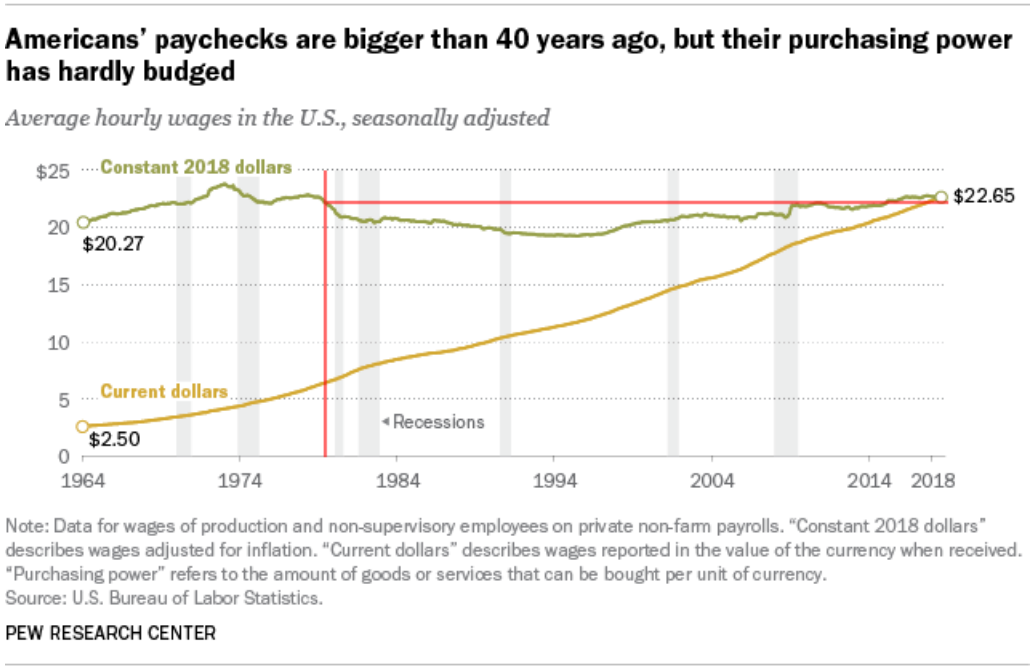
So the two main differences are that the CPI is a measure of purchasing power, while the PCE is a measure of actual spending. The CPI compares wages to the average cost of things people buy based on household receipts, while the PCE compares wages to the official sticker price of the goods and services people actually receive.

In our opinion, the CPI is a better indicator of "economic power" than the PCE - the PCE may be a better indicator of standard of living not controlling for debt, but even that is dubious because it assumes the "real value of things" to be what businesses say their products are worth; not what the typical person actually pays for something.

Another factor that neither of these measure take into account is "non-market goods." For example, a family with a stay-at-home wife has maid services, lawn services, daycare services, a

cook, and if she homeschools the kids - a private tutor. If one was to pay for these things privately, that would register in the economy as more economic activity. This is always something to consider when looking at "economic growth" - it may just be things that used to be free and not part of the "economy" per se have become financialized.

All those caveats aside, we can look at some measure of changes in wages compared to the CPI and PCE. Pew looked at the change in US wages from 1964 to 2018:

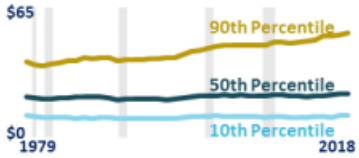
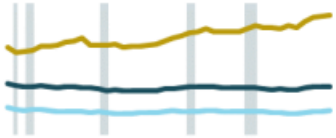
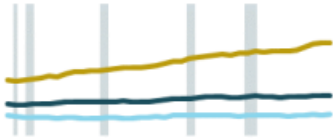
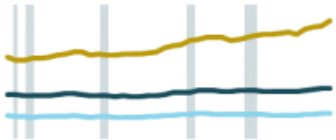
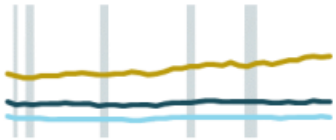
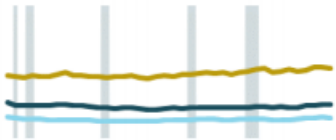
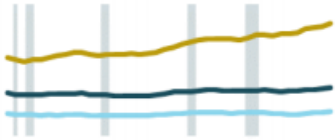


I added the red lines to mark the year 1979. Those red lines are not on the original image from Pew. This is because a report by the Congressional Research Service[31] starts their analysis in 1979. By 2018, real wages have slightly edged out 1979, but failed to edge out the peak around 1972.

According to the Congressional Research Service, which uses a form of the CPI, real wages have gone up 6.1% for the 50th percentile, 1.6% for the 10th percentile, and 37.6% for the 90th percentile:

[31] https://fas.org/sgp/crs/misc/R45090.pdf

Table 1. Real Wage Trends over 1979-2018, by Selected Demographic Characteristics

Demographic	Real Wage Trends	Cumulative % Change in Real Wages		
	Shaded Bars = Recessions	*10th percentile*	*50th percentile*	*90th percentile*
Overall		1.6%	6.1%	37.6%
Men		-13.3%	-5.1%	36.4%
Women		4.8%	25.7%	66.7%
White (Non-Hispanic)		8.2%	13.2%	45.6%
Black (Non-Hispanic)		-0.3%	1.2%	28.5%
Hispanic		-3.7%	-4.6%	11.4%
Non-Hispanic		6.7%	10.1%	42.7%

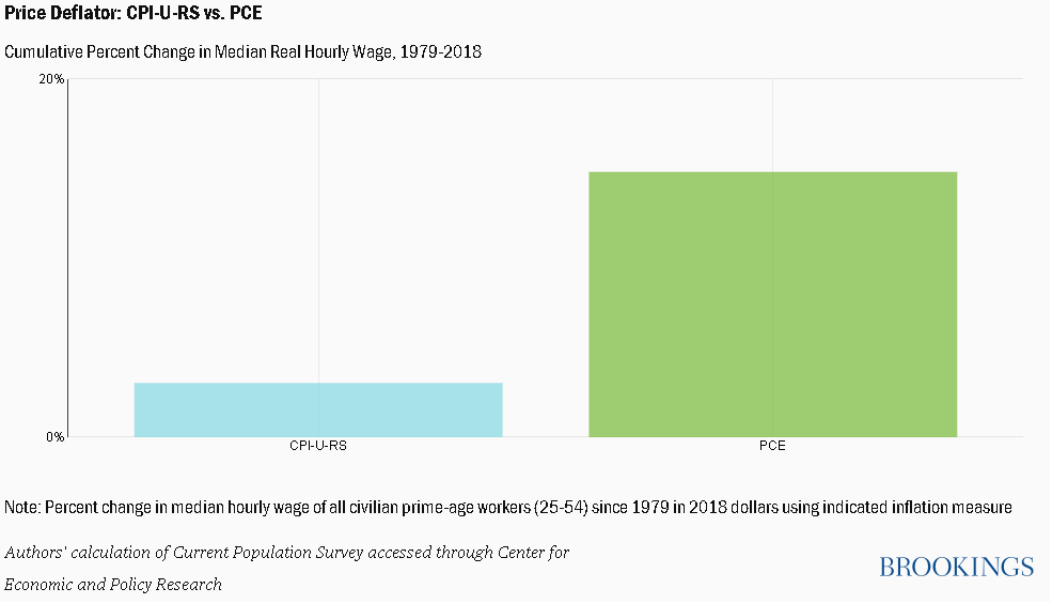
However, the gains are being driven entirely by female wage growth; male wages have declined. If women in the past were more likely to work part time before finding a man and then having kids and becoming a stay-at-home mom, this change is suddenly a lot less impressive from a purely economic standpoint. This is not to cast some moral judgement on women entering formal economic work as opposed to being in the informal economy. It is simply to say, as a matter of fact, that men don't make as much as they used to in 1979 according to the CPI, and women don't make as

much as men did in 1979.

The gains in female wages may be slightly - and we should emphasize slightly - greater in absolute terms than the decline in male wages since 1979. But this is a function of women taking work more seriously, to a degree men always have. And the fact is that these people - men and women - are earning less despite trying as hard as the people in 1979 did - who were moreso men. That's the stagnation.

This is also not counting the "unpaid" labor that women were more likely to do in 1979 and before, however much that adds up to. While this doesn't factor into econometric wage measurements, it does impact real standard of living. If we were somehow able to factor in the loss of the informal economy, I believe we would see a substantial decline in "real wages plus informal benefits" has occurred.

So what is happening is people are getting paid less for the same effort - the fact that more women are taking work more seriously merely masks this problem. And all of the informal benefits of women not being in the formal workforce (whatever your moral opinions on that are) are lessened to the extent women are in the workforce.



In terms of what the household gets - one man could work, have job security, and have about as high real wages as today. In addition, the woman works off the books, and the household has the services of a maid, a cook, a nanny, a daycare, a groundskeeper, janitor and butler. And this is what is really causing people - especially men - to think that the economy is stagnating.

It's also interesting to note that the black-white wage gap has increased from 1979. This calls into question how important any net anti-black racial discrimination was in 1979, if it even existed.

But that digression aside, the takeaway is that wages have, on the surface level, been stagnant since at least 1972. Incidentally that is the last year a man walked on the moon. And the eradication of the informal economy, or the financialization of things that used to be "free," should make one even more skeptical regarding claims there has not been stagnation.

That said, we Brookings looked at how the picture changes when, instead of using the CPI, we

use the PCE index.

So the PCE - Personal Consumption Expenditures - shows much greater wage growth than does the CPI. But keep in mind, even the PCE numbers say that real wages have grown 15% over 29 years. Or 0.517% per year.

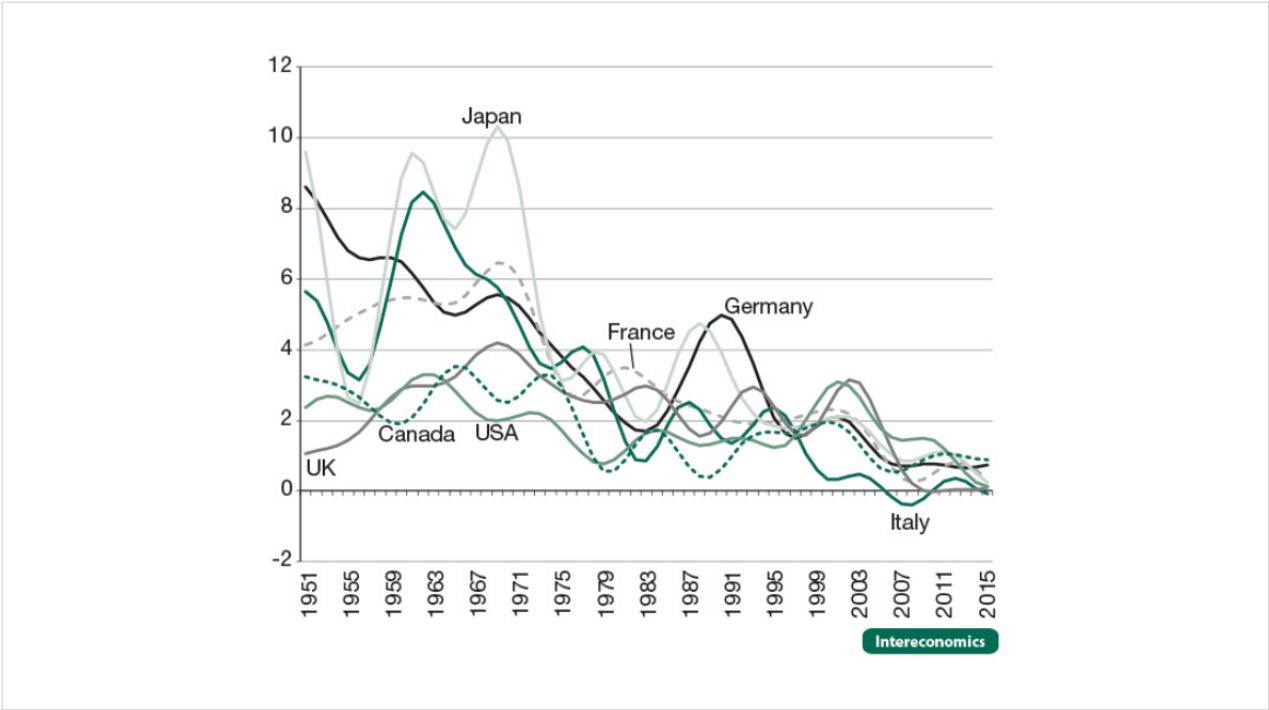
The last thing to look at on Economic stagnation is "total factor productivity growth". Total Factor Productivity is the ratio of the value of capital and labor and the value of outputs. It is a rough indicator of how efficient a economy is with what is put into it.

The paper "Global Productivity Slowdown: Diagnosis, Causes and Remedies"[32]. looks at the changes in Total Factor Productivity GROWTH from 1951 to 2015 in France, the US, the UK, Germany, Italy, Spain, Canada and Japan:

Figure 1

Labour productivity trends in the G7

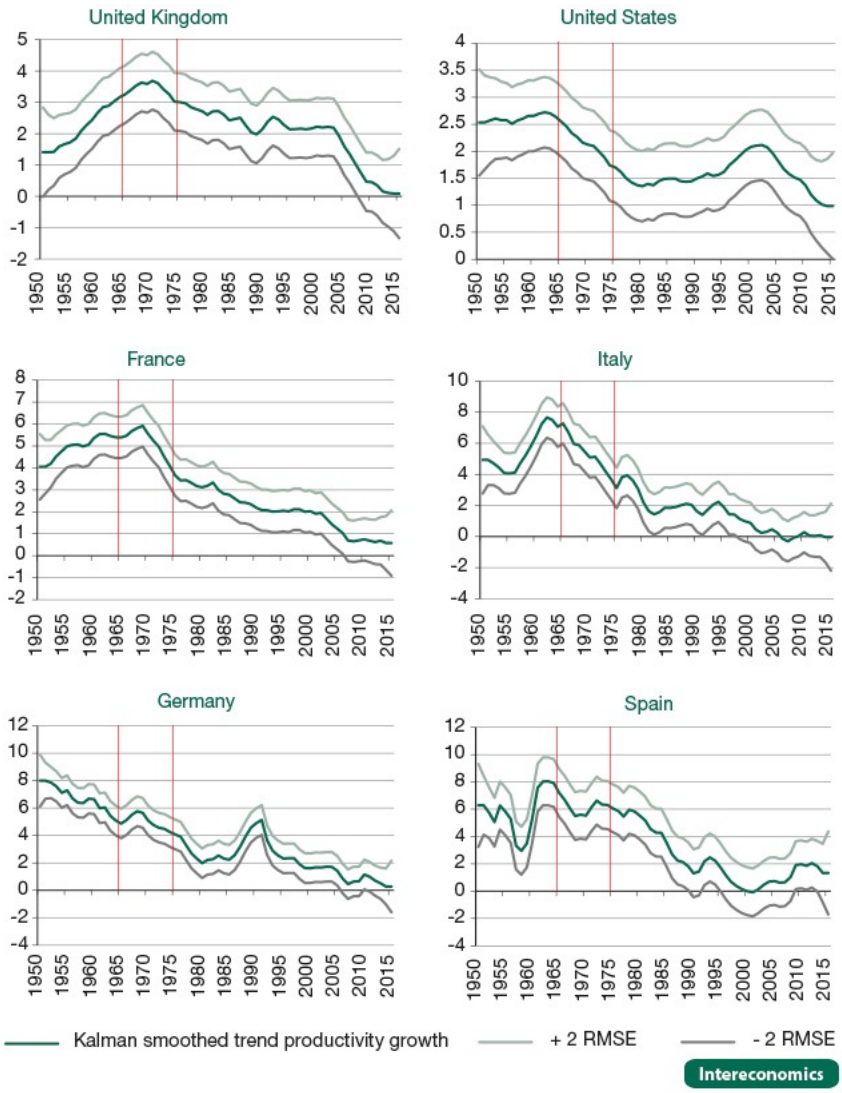
in %



Source: The Conference Board Total Economy Database: Output, Labor and Productivity, 1950-2015. We calculated trends using the asymmetric version of the Christiano-Fitzgerald band pass filter under the assumption that the original series are integrated of order 1. See L.J. Christian, T.J. Fitzgerald: The Band Pass Filter, in: International Economic Review, Vol. 44, No. 2, 2003, pp. 435-465.

In general, the trend is down over time, from a peak of all of these countries' TFP growth being between 1965 and 1973. The paper breaks down six individual countries:

[32]https://www.intereconomics.eu/contents/year/2017/number/1/article/the-global-productivity-slowdown-diagnosis-causes-and-remedies.html



The red lines were added to more clearly show the decade 1965-1975. Those red lines are not in the original images.

For all of these countries, that decade was either a period of steep decline, or the beginning of the decline. This is not to be a referendum on the validity of Total Factor Productivity as a measure, but certainly TFP measures something, and that something is generally going down. And this decline began, among these 8 countries, roughly between the years 1965 and 1975 or shortly thereafter.

5.2 Technological Stagnation

The question of "technological stagnation" is harder to answer because inventions don't have science points. How many science points is an iPad worth?

One thing that we recommend you look into are predictions of the future people made in the past. Today people have a habit of ha-haing such notions as naive. But think of something that an

80 year old woman in 1969 had seen. She had likely seen her home go from not having electricity to being "electrified," perhaps witnessed the introduction of running water in her own home and certainly in someone else's. She saw the growth of trains, the rise of the automobile and the end of horse-carriages on major urban streets. She saw the rise of the telephone, the moving picture, then the airplane, commercial air travel, the jet engine and faster-than-sound aircraft. She saw the exploration of the ocean floors, offshore oil rigs, the introduction of plastics, the rise of computers, antibiotics and medical marvels that rendered once-dangerous diseases trivial, and toward the end of her days she saw a man walking on the moon.

While difficult to quantify, to say there has been no visible stagnation would be lying to oneself. Outside of the "world of bits," computer technology, how different are things really from 1970? That's not to say there are *no* differences - that cars haven't improved, that ovens and microwaves and refrigerators aren't better. But look around your home. Take away any screens. Outside of issues of design, how do you know you're not in 1970?

You'll probably find some things. Maybe the electric plugs are a little better, the refrigerator or washing machine a little more reliable. Eric Weinstein had a interesting quote to those who bring up their iPad or any other invention in the world of bits:

> "Of course your iPad is amazing - that's all that's left of your once limitless future."

Maybe it's just that the low-lying fruit has been picked? But lets establish that there has in fact been technological stagnation first. The paper "Are Ideas Getting Harder to Find"[33] comes up with quantification of output of various fields versus measure of how much research effort is put into them. Of course the focus is on that which is most quantifiable - trying find some objective hook into an inherently subjective question.

First, he looks at agriculture. Agricultural yield growth has been declining in the four crops he looked at since 1960, with the arguable exception of Cotton which had a yield growth spike in the 1980s. But even Cotton yield growth has been in decline since the 1980s. On this, one might say (somewhat literally) that we are reaching diminishing returns since the low-hanging "fruit" has been plucked. Fair enough, there are, in its first instance, all manner of explanations.

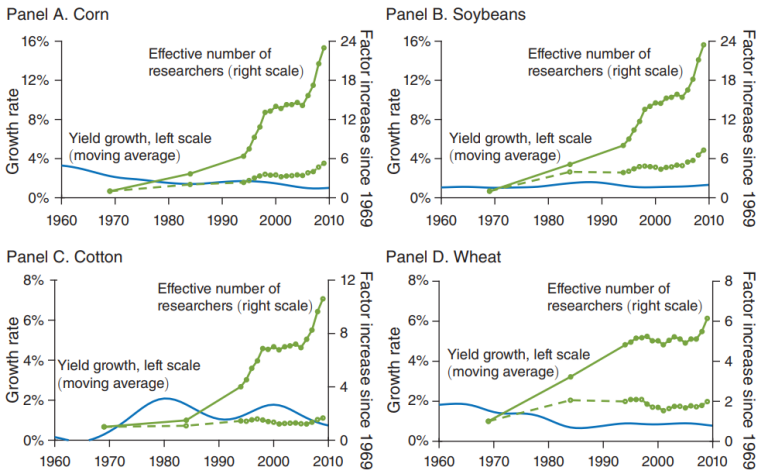


FIGURE 6. YIELD GROWTH AND RESEARCH EFFORT BY CROP

[33] https://web.stanford.edu/ chadj/IdeaPF.pdf

The paper also looks at the decline in years of life gained from clinical trials and papers published. The relation between life extension and number of papers published seems small, but the relation between life extension and clinical trials is dropping precipitously. Again, it could just be that the low-hanging fruit has been picked, or it could be something else, or perhaps a combination.

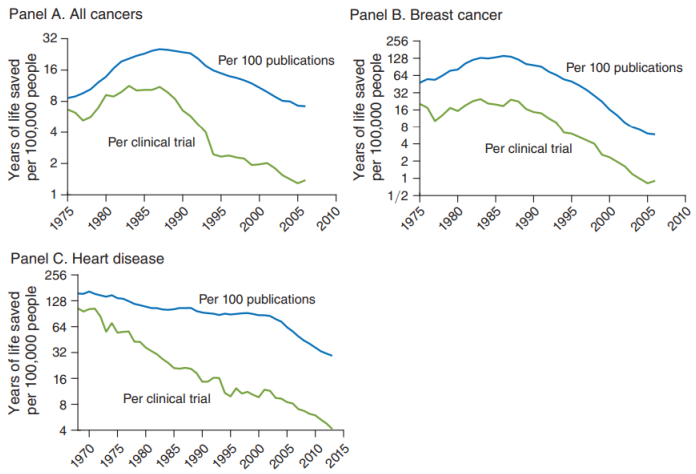


FIGURE 9. RESEARCH PRODUCTIVITY FOR MEDICAL RESEARCH

Note: Research productivity is computed as the ratio of years of life saved to the number of publications.

The paper also looked at changes in "Total Factor Productivity" in the United States. The TFP growth estimate they use is for the next 5 years at any point in the graph, which is why it looks different from the previous graphs on total factor productivity in the United States. But the takeaway is that while there are over 100 times as many researchers, averaged total factor productivity growth is less than it was in 1950, inching it's way down to zero.

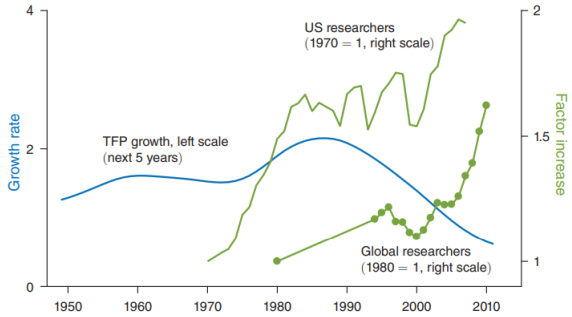


FIGURE 5. TFP GROWTH AND RESEARCH EFFORT IN AGRICULTURE

Notes: The effective number of researchers is measured by deflating nominal R&D expenditures by the average wage of high-skilled workers. Both TFP growth and US R&D spending (public and private) for the agriculture sector as a whole are taken from the US Department of Agriculture Economic Research Service (2018a,b). The TFP series is smoothed with an HP filter. Global R&D spending for agriculture is taken from Fuglie et al. (2011), Beintema et al. (2012), and Pardey et al. (2016).

They then look at changes in research productivity for manufacturing, looking at how much research was put into the field on an index of papers and researchers and money, and how much this was correlated with improvements in measures of manufacturing efficiency. The Average growth

is annual. Meaning that every year, research is getting somewhere between 4.9% and 8.1% **less** productive per year.

TABLE 6—CENSUS OF MANUFACTURING RESULTS, ACROSS TWO DECADES (1992–2002, 2002–2012)

	Effective research		Research productivity	
Case	Factor increase	Avg. growth (%)	Factor decrease	Avg. growth (%)
1. Benchmark	1.2	1.6	2.2	−7.8
2. Winsorize $g < 0.01$	1.2	1.6	1.9	−6.0
3. Winsorize top/bottom	1.2	1.6	1.7	−4.9
4. Unweighted	1.0	0.0	1.9	−8.1
5. Research = scientists	1.3	2.3	2.3	−6.0

Notes: Research productivity is the ten-year DHS growth in real sales divided by mean R&D spending, deflated by the skilled wage, over those ten years. Research productivity growth is then calculated as the percent change in research productivity compared to ten years earlier. In row 2, idea output (sales growth) is winsorized from below at 1 percent. In row 3, idea output (sales growth) is winsorized from below at 1 percent and from above such that an equal number of firms are winsorized in each tail. In row 4, the mean is unweighted. In row 5, the denominator in research productivity is the number of scientists and engineers. In rows 1 to 3, the mean of the growth rate of R&D is weighted by mean R&D over the past 20 years. In row 5, the mean of the growth rate of scientists and engineers is weighted by mean R&D over the past 20 years. *Factor decrease* is calculated as $1/(1 - \text{mean})$ where mean is the mean of the research productivity growth weighted by the average R&D spending over the past 20 years. *Average growth* is calculated as $1 - (1 - \text{mean})^{1/10}$ where mean is the mean of research productivity growth weighted by the average R&D spending over the past 20 years. The sample includes 1,300 firms and 2,700 observations for all cells.

The authors summarize their results in table A1:

TABLE 7—SUMMARY OF THE EVIDENCE ON RESEARCH PRODUCTIVITY

Scope	Time period	Average annual growth rate (%)	Half-life (years)	Dynamic diminishing returns, β
Aggregate economy	1930–2015	−5.1	14	3.1
Moore's Law	1971–2014	−6.8	10	0.2
Semiconductor TFP growth	1975–2011	−5.6	12	0.4
Agriculture, US R&D	1970–2007	−3.7	19	2.2
Agriculture, global R&D	1980–2010	−5.5	13	3.3
Corn, version 1	1969–2009	−9.9	7	7.2
Corn, version 2	1969–2009	−6.2	11	4.5
Soybeans, version 1	1969–2009	−7.3	9	6.3
Soybeans, version 2	1969–2009	−4.4	16	3.8
Cotton, version 1	1969–2009	−3.4	21	2.5
Cotton, version 2	1969–2009	+1.3	−55	−0.9
Wheat, version 1	1969–2009	−6.1	11	6.8
Wheat, version 2	1969–2009	−3.3	21	3.7
New molecular entities	1970–2015	−3.5	20	…
Cancer (all), publications	1975–2006	−0.6	116	…
Cancer (all), trials	1975–2006	−5.7	12	…
Breast cancer, publications	1975–2006	−6.1	11	…
Breast cancer, trials	1975–2006	−10.1	7	…
Heart disease, publications	1968–2011	−3.7	19	…
Heart disease, trials	1968–2011	−7.2	10	…
Compustat, sales	3 decades	−11.1	6	1.1
Compustat, market cap	3 decades	−9.2	8	0.9
Compustat, employment	3 decades	−14.5	5	1.8
Compustat, sales/employment	3 decades	−4.5	15	1.1
Census of Manufacturing	1992–2012	−7.8	9	…

Notes: The growth rates of research productivity are taken from other tables in this paper. The half-life is the number of years it takes for research productivity to fall in half at this growth rate. The last column reports the extent of dynamic diminishing returns in producing exponential growth, according to equation (17). This measure is only reported for cases in which the idea output measure is an exponential growth rate (i.e., not for the health technologies, where units would matter).

What's interesting here is that even in the world of bits research productivity is collapsing.

Another way to look at this is by number of patents. Again, no metric is perfect - some patents could simply be better than others. We may just be getting higher quality patents than in the

past. That's a possibility. We could also be getting "worse" quality patents in the sense that they are more a bunch of minor improvements and not really groundbreaking as they used to be. Since we can speculate that the "science points per patent" could either be going up or down, looking at the raw number of patents should be treated as some generalized index of innovation *more* or *less*.

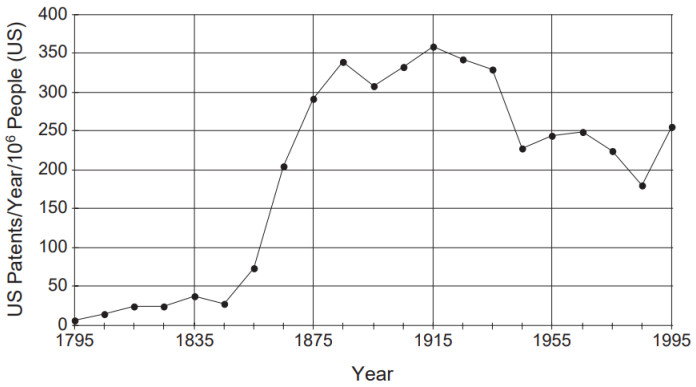


Fig. 3. Rate of invention. Points are an average over 10 years with the last point covering the period from 1990 to 1999.

This is from Huebner 2005[34], looking at the number of patents per 1 million people per year. He points to a spike up to 1995, where his analysis stopped.

Huebner also looks at the rate of "significant scientific events" per year per 1 million people.

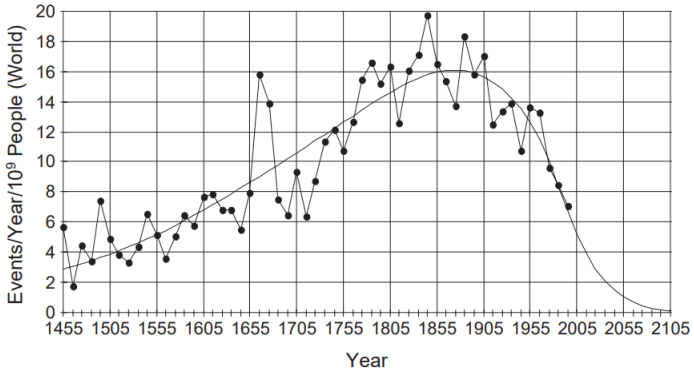


Fig. 1. Rate of innovation since the end of the Dark Ages. Points are an average over 10 years with the last point covering the period from 1990 to 1999. The smooth curve is a least squares fit of a modified Gaussian distribution to the data.

Again, Huebner has a definition for what is a "significant event." It could underplay modern significant events because those events are more difficult to see, they're more complex. Perhaps. Or it could have a recency bias where smaller advances in the past get buried and lost in time, and that

[34]https://www.sciencedirect.com/science/article/abs/pii/S0040162505000235

in the past there was a higher threshold for "signifcant." Again, we can speculate ways in which past innovations are underplayed or overplayed, and which modern innovations are underplayed or overplayed, and since we don't know which is the greater effect, just going with the raw number of Huebner-defined "significant events" is the best guesstimate for the overall "science points" scored per million people per year.

Now this graph gives some perspective and begs a question different than what we normally ask. Which is not "why such slowdown after 1960?", but moreso "why the rise up to 1960?" i.e. - it is the period from 1755 to 1905 that pops out as anomalous on this graph. Though other measures may put the years 1900 to 1970 as the big rise. Either way, in the broad arc of history, some time between 1750 and 1970 was the period of tremendous scientific innovation, and the way to look at it may be "what was so special about this period" as opposed to the more myopic question of "why is science declining *now*?"

So this great stagnation appears to be happening. Each line of evidence is flawed, but multiple lines are saying the same thing.

Why?

The first explanation most leap to is "low-hanging fruit." This is certainly a possibility. It could be some fields are in fact getting more *natively* difficult to advance in. That is, the nature of the problems being solved are becoming more complex.

It's also possible that some fields are getting *natively* easier to advance in. That is, once you make a few key breakthroughs, it becomes easier to advance in certain fields. And that the apparent slowdown is a result of institutional failure in spite of field becoming natively easier to advance in.

So it could be one, the other, both, sometimes one or the other depending on the field or the particular problem. The "low-hanging fruit" idea is not some default assumption or starting point anymore than the idea that science is natively getting easier is the default starting point.

There has been some economic stagnation. Wage stagnation cannot account for the improvement in quality of products, but also cannot account for the loss of informal labor - things that used to not be measured. And all measures of scientific advancement - however flawed each one is individually - all paint the same picture.

There are several factors that point to the "institutionalization" or "corporatization" of academia being the cause of this stagnation, however.

5.3 Stagnation occurred in multiple fields at the same time

While looking through a glass darkly, Wikipedia has a list of the "paradigm shifts of the natural sciences." These paradigm shifts, instead of accelerating with the growing population, seem to have stopped at 1985 at the latest, 1974 at the earliest, according to this subjective list. Now this list is subjective of course, but when compiling it, the editors were not trying to prove some great stagnation. They just went with what they thought were the big paradigm shifts.

Natural sciences [edit]

Some of the "classical cases" of Kuhnian paradigm shifts in science are:

- 1543 – The transition in cosmology from a Ptolemaic cosmology to a Copernican one.[12]
- 1543 – The acceptance of the work of Andreas Vesalius, whose work *De humani corporis fabrica* corrected the numerous errors in the previously-held system created by Galen.[13]
- 1687 – The transition in mechanics from Aristotelian mechanics to classical mechanics.[14]
- 1783 – The acceptance of Lavoisier's theory of chemical reactions and combustion in place of phlogiston theory, known as the chemical revolution.[15][16]
- The transition in optics from geometrical optics to physical optics with Augustin-Jean Fresnel's wave theory.[17]
- 1826 – The discovery of hyperbolic geometry.[18]
- 1859 – The revolution in evolution from goal-directed change to Charles Darwin's natural selection.[19]
- 1880 - The germ theory of disease began overtaking Galen's miasma theory.
- 1905 – The development of quantum mechanics, which replaced classical mechanics at microscopic scales.[20]
- 1887 to 1905 – The transition from the luminiferous aether present in space to electromagnetic radiation in spacetime.[21]
- 1919 – The transition between the worldview of Newtonian gravity and general relativity.
- 1964 - The discovery of cosmic microwave background radiation leads to the big bang theory being accepted over the steady state theory in cosmology.
- 1965 - The acceptance of plate tectonics as the explanation for large-scale geologic changes.
- 1974 - The November Revolution, with the discovery of the J/psi meson, and the acceptance of the existence of quarks and the Standard Model of particle physics.
- 1960 to 1985 - The acceptance of the ubiquity of nonlinear dynamical systems as promoted by chaos theory, instead of a laplacian world-view of deterministic predictability. [22]

These paradigm shifts cover a lot of fields. A great coincidence that they would all sputter out at the same time.

Sean last wrote in a paper which at the time of this writing is unpublished, more of his subjective accounts with people:

> "In psychology, the typical list given for this sort of thing would stop with Humanistic psychology and the Cognitive revolution, thus ending in the 1950s. In economics, a typical list would probably end with Keynesianism, Monetarism, and maybe a few other theoretical developments but would probably not extend past the 1970s. In statistics, such a list would end with things like pathway analysis, significance testing, and meta-analysis, and so again would end in the mid 20th century. In philosophy, a list of paradigm shifts would probably end with post-modernism, existentialism, and analytic philosophy, all of which occurred prior to 1970. People I've spoken to have told me that the same is true of fields I know less about, such as linguistics and history.
>
> Of course, this is not to say that there's been no progress in these fields. But the progress that has happened has either been incremental progress where details are added to pre-existing paradigms, or potential new paradigms that fail to gain widespread acceptance (e.g. evolutionary psychology)."
>
> - Sean Last, unpublished

While we cannot prove this to you, these conversations are what led Sean to look into the "great stagnation"; he was not fishing for confirmation of this at the time.

This is the first problem with the "low-hanging fruit" hypothesis: that all of these advances would sputter out at the same time. All the low-hanging fruit grabbed all at once? Or maybe the fruit was always at intermediate height and we just got shorter?

This is somewhat muddled by the fact that advances have apparently continued to some degree in the world of bits. But even on that the gains haven't been as rapid as are lauded, and it's an exception which ties into another reason for stagnation.

5.4 This stagnation occurred with the rise of the journal system and the "corporatization" or "institutionalization" of academia.

The period of general scientific stagnation corresponded with the rise of "peer review", or the journal system. Now even "the journal system", while more accurate, is a bit of a misnomer since Journals became popular long before the rise of review boards in those journals, and so what actually happened from 1920-1970 was the rise of review boards for articles within those journals.

The paper "Scientific Autonomy, Public Accountability, and the Rise of 'Peer Review' in the Cold War United States"[35] describes the rise of review boards for articles in journals in the west:

> "In the interwar period, a broader push toward standardization during the Progressive Era had a tremendous impact on scientific practice, particularly in the United States. The quest for standardization seems to have been one impetus that influenced the development of increasingly formal refereeing procedures at British and American scientific societies. Refereeing procedures at the American physics journal Physical Review, for example, became much more standardized during the 1920s and 1930s. Referees who had once written free-form letters sharing their general impressions of Physical Review submissions were now asked to fill out forms assessing a paper's suitability according to a predetermined list of criteria. However, most papers accepted for Physical Review never went out to referees at all; the editor accepted most papers on his own authority, consulting referees only when he thought he might want to reject a paper. It was not until the 1960s that all Physical Review papers were sent out for external referee opinions.
>
> ...
>
> Other prominent English-language journals adopted systematic external refereeing even later. The New England Journal of Medicine began having two outside reviewers consider all potentially acceptable papers in the late 1960s. Nature began employing referees for every paper it published only in 1973. Well into the 1970s, the British medical journal the Lancet relied heavily on editorial judgment, with editors accepting or rejecting up to 90 percent of submissions themselves.
>
> ...
>
> Many funding bodies had unsystematic or internal review processes that placed heavy responsibility in the hands of organization employees. Private funding bodies such as the Rockefeller Foundation, for example, generally left funding decisions in the hands of trusted middle managers well into the postwar period, awarding money via what Robert Kohler has described as a 'patronage system.' The same was true for many publicly funded grant organizations. The German Research Foundation, created in 1920 and initially called the Emergency Association for German Science, deliberately chose to rely on a small number of elite scientists for opinions on grant proposals, and much of the evaluation focused on the personal qualities of the applicants. Well into the twentieth century, a single three-man committee evaluated all applications for the Royal Society of London's Government Grants; though all were invited to apply, the process awarded those grants almost exclusively to Fellows of the Royal Society."

[35] https://sci-hub.tw/10.1086/700070

Based on this account, the beginning of the great stagnation overlaps with the rise of formalized review boards for papers. Something which we have shown the effects of at length earlier. In addition, these institutional changes have coincided with an explosion of scientific papers. This is because any academic resume is now expected to have a long list of publications.

5.5 There are intuitive causal ways that the changes in "institutional science" can cause a decline

From Jonathan Katz, professor at Physics at Washington University, generally recommends talented individuals to *not* become scientists. He describes why:

> "Suppose you do eventually obtain a permanent job, perhaps a tenured professorship. The struggle for a job is now replaced by a struggle for grant support, and again there is a glut of scientists. Now you spend your time writing proposals rather than doing research. Worse, because your proposals are judged by your competitors you cannot follow your curiosity, but must spend your effort and talents on anticipating and deflecting criticism rather than on solving the important scientific problems. They're not the same thing: you cannot put your past successes in a proposal, because they are finished work, and your new ideas, however original and clever, are still unproven. It is proverbial that original ideas are the kiss of death for a proposal; because they have not yet been proved to work (after all, that is what you are proposing to do) they can be, and will be, rated poorly. Having achieved the promised land, you find that it is not what you wanted after all."

David Graeber, Anthropologist from the University of Chicago, wrote of the change in University Structure from 1972 to 2012:

> "What has changed is the bureaucratic culture. The increasing interpenetration of government, university, and private firms has led everyone to adopt the language, sensibilities, and organizational forms that originated in the corporate world. Although this might have helped in creating marketable products, since that is what corporate bureaucracies are designed to do, in terms of fostering original research, the results have been catastrophic.
>
> My own knowledge comes from universities, both in the United States and Britain. In both countries, the last thirty years have seen a veritable explosion of the proportion of working hours spent on administrative tasks at the expense of pretty much everything else. In my own university, for instance, we have more administrators than faculty members, and the faculty members, too, are expected to spend at least as much time on administration as on teaching and research combined. The same is true, more or less, at universities worldwide.
>
> The growth of administrative work has directly resulted from introducing corporate management techniques. Invariably, these are justified as ways of increasing efficiency and introducing competition at every level. What they end up meaning in practice is that everyone winds up spending most of their time trying to sell things: grant proposals; book proposals; assessments of students' jobs and grant applications; assessments of our colleagues; prospectuses for new interdisciplinary majors; institutes; conference

> workshops; universities themselves (which have now become brands to be marketed to prospective students or contributors); and so on.
>
> As marketing overwhelms university life, it generates documents about fostering imagination and creativity that might just as well have been designed to strangle imagination and creativity in the cradle. No major new works of social theory have emerged in the United States in the last thirty years. We have been reduced to the equivalent of medieval scholastics, writing endless annotations of French theory from the seventies, despite the guilty awareness that if new incarnations of Gilles Deleuze, Michel Foucault, or Pierre Bourdieu were to appear in the academy today, we would deny them tenure.
>
> There was a time when academia was society's refuge for the eccentric, brilliant, and impractical. No longer. It is now the domain of professional self-marketers. As a result, in one of the most bizarre fits of social self-destructiveness in history, we seem to have decided we have no place for our eccentric, brilliant, and impractical citizens. Most languish in their mothers' basements, at best making the occasional, acute intervention on the Internet."

Something to consider is how Isaac Newton wrote the *Principia Mathematica*. Newton was born in 1643 and attended Trinity College in 1661 at age 18. Newton was regarded as a mediocre student. In 1665 the University shut down due to the plague, and Newton went home and, in private, wrote works on calculus, optics and the law of gravitation. Newton published very little between 1666 and 1687, when he published the *Principia* and laid out Newtonian physics.

Imagine if Newton was in a modern university system. As a mediocre student, he wouldn't have gotten a professorship. If he did, he would have gotten it by around age 35. In the interim, he would have had to work of projects for a senior professor, and would have had to publish articles on a regular basis, ideally in prestigious journals, in order to remain competitive for a position.

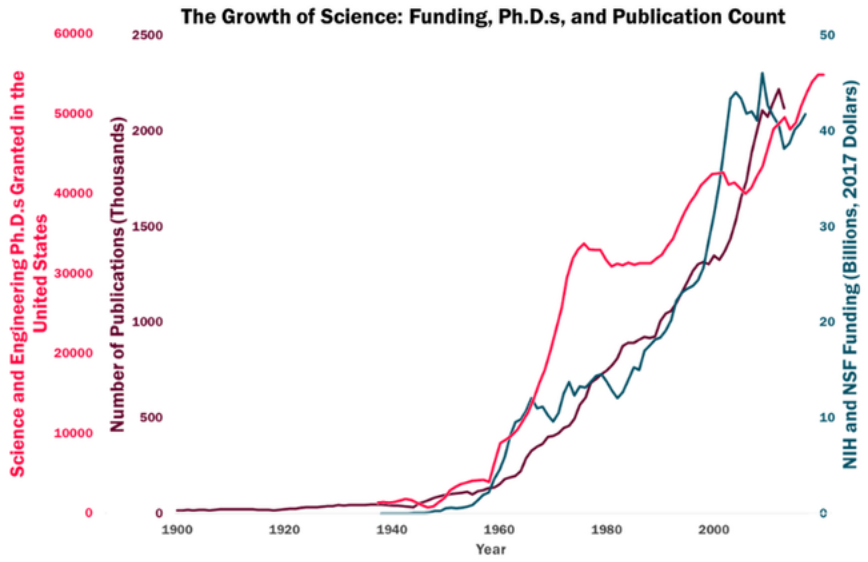
In reality, what happened is Newton published a few articles on calculus and optics and laws of motion, never went through a formal "peer review" process, and really only published on big thing at the end. Now perhaps science is so complex that it's impossible for a lone eccentric genius to contribute anything, perhaps being a key term - because we'd never know since it's impossible for someone like Newton to ever do what he did today. In today's system, Newton would be a mediocre student and that would be the end of his story.

He might have some crazy ideas that he'd post on a geocities page somewhere, along with pages on astrology and alchemy (which Newton also believed in – in fact, alchemy was his primary interest). His interest in those topics would discount his heterodox work on physics. Anyone who saw his website would just think he's one among many cranks, then one day his geocities account would expire, and that would be it.

5.6 Explosion in the number of PhDs

From the article "Science Is Getting Less Bang for Its Buck"[36], Collison and Nielsen looked at the number of PhDs, publications and NIH and NSF funding from 1900 to 2017, and show the skyrocketing increase over these years:

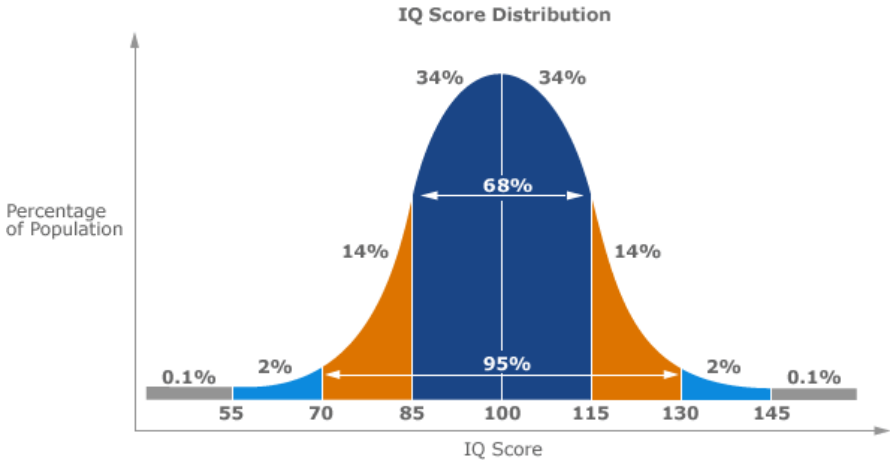
[36] https://www.theatlantic.com/science/archive/2018/11/diminishing-returns-science/575665/



This has predictably coincided with a collapse in the productivity of research, as well as an absolute decline in novel breakthroughs.

5.7 A Decline in "g"

Now we're going to say some things here that are controversial and will latch onto concepts that will be established later in this seminar. Which is that there has been a general decline in "g", or the "general intelligence factor." This is not the same as "IQ", though the two are related. If you are sufficiently socialized, you will have heard that nominal IQs have been rising. And this is documented from around 1890 to sometime in the late 1990s, depending on location, and was probably true well before it was documented. However, recently there have been declines in what are called "raw IQ scores." The way an IQ score works is that your score is actually a rank. A score of 100 means you are at the 50th percentile, and 115 means you're 1 standard deviation above the 50th percentile. This is visualized in a bell curve:



So when "raw scores" go up or down, the average and median IQ is still, by construction, 100. However, what that 100 means can change. So raw scores were going up from it's earliest measurements around the turn of the century to sometime in the 1990s. This rise in scores was coined by Charles Murray "The Flynn Effect." Since then the raw scores have been going down.

The paper "The Negative Flynn Effect: A systematic literature review"[37] documents the recent declines in some countries:

Table 1
Negative Flynn Effect per country.

Country	Age	Test	Years	Type	IQ (decline per decade)	Reference
Norway	18–19	General ability	1996–2002	All conscripts in every year	0.38	Sundet al., 2004
Denmark	18–19	Borge Priene's Prove	1998–2003/4	All conscripts in every year	2.70	Teasdale & Owen, 2008
Britain	11–12	Piagetian	1975/2003	10,023 over 5 cohorts: 1975, 2000, 2001, 2002, 2003 (each cohort roughly equal in size)	4.30	Shayer & Ginsburg, 2007
Britain	13–14	Piagetian	1976/2006	2006: N 446, 2007: N 357 (total: 793)	2.50	Shayer & Ginsburg, 2009
Netherlands	Adults	GATB	1975/2005	Meta-analysis	1.35	Woodley & Meisenberg, 2013
Finland	18–19	Peruskoe	1998–2009	All conscripts 1998–2001 and 2008/9	2.0	Dutton & Lynn, 2013; Koivunen, 2007
France	Adults	WAIS III & IV	1999/2008–9	Two representative groups of 79	3.8	Dutton & Lynn, 2015
Estonia	18–19	Raven SPM	2001/2005/2012	Representative student sample: 2001: 573, 2005: 417, 2012: 338	8.4	Korgesaar, 2013

Just as the positive Flynn Effect wasn't uniform around the world or, in this case, even within Europe, the negative Flynn Effect also does not appear to be uniform, hitting different regions, professions, age groups, differently.

This is just nominal IQ, not "g" itself. To put it in simple terms, "IQ""is the measurement, while "g" is the underlying capacity. And this underlying capacity can be inferred, and was shown to be predictively valid (i.e. was operationally validated). We'll talk about the g factor later in this seminar.

Given that there's no time in this section to describe the methods for inferring "g", our hope is that you hold this explanation for the decline of scientific productivity in a kind of "super-position." That it exists as a possibility, as yet unproven as you likely don't understand and - based on your place and time in history - are predisposed to disbelieving intelligence tests and have likely heard compelling-sounding arguments against them.

That said, using reaction time tests which are used as a measure of "g," we can see a decline in "g" from 1889, from the paper "Were the Victorians cleverer than us? The decline in general intelligence estimated from a meta-analysis of the slowing of simple reaction time"[38]:

> "The difference between the meta-regression trend-weighted present (2004) simple RT mean (275.47 ms) and the trendweighted 1889 mean (194.06 ms) is 81.41 ms."

Unfortunately, the data is spotty going back this far, and relies on reaction-time data. Here were the studies they used:

[37] https://sci-hub.se/10.1016/j.intell.2016.10.002
[38] https://sci-hub.tw/10.1016/j.intell.2013.04.006

Table 1
14 simple RT studies used in Silverman (2010) and Thompson (1903) along with 16 simple RT means, sample sizes, collection/publication year and references.

Testing year and country	Males (*N*)	Females (*N*)	Sample size weighted mean (total *N*)	Reference
1889[a] (1884–1893) (UK)	183 (2522)	187.9 (888)	184.3 (3410)	Galton's data in Johnson et al. (1985)
1894.5[a] (1889–1900) (USA)	199 (24)	217 (25)	208 (49)	Thompson (1903)
1941 (USA)	197 (47)	n.a	197 (47)	Seashore, Starmann, Kendall, and Helmick (1941)
1941 (USA)	203 (47)	n.a	203 (47)	Seashore et al. (1941)
1945 (UK)	286 (76)	n.a	286 (76)	Forbes (1945)
1970 (Canada)	236 (40)	263 (40)	249.5 (80)	Lefcourt and Siegel (1970)
1990 (Finland)	199 (20)	n.a	199 (20)	Taimela (1991)
1987 (Finland)	183 (20)	n.a	183 (20)	Taimela, Kujala, and Osterman (1991)
1993 (USA)	260 (80)	285 (140)	275.9 (220)	Anger et al. (1993)
1993 (USA)	250 (73)	280 (163)	270.7 (236)	Anger et al. (1993)
1999 (UK)	306 (64)	n.a	306 (64)	Smith et al. (1999)
2002 (UK)	324 (24)	n.a	324 (24)	Brice and Smith (2002)
1999.5[a] (1999–2000) (Australia)	214 (1163)	224 (1241)	219.5 (2404)	Jorm, Anstey, Christensen, and Rodgers (2004)
2004 (Canada)	253 (171)	268 (198)	261 (369)	Reed, Vernon, and Johnson (2004)
1987.5 (1987–1988) (UK)	295 (254.5)[b]	306 (288.5)[b]	300.3 (543)	Deary and Der (2005a)
1984.5 (1984–1985) (UK)	300 (834)	318 (1023)	309.6 (1857)	Der and Deary (2006)

Additional. We went back to Johnson et al. (1985) and cross-referenced it with Silverman (2010). The total *N* for females should be 888 rather than 302. We changed the above *N* to reflect the correct females sample size.
[a] When a range of years is given the average is taken.
[b] In these studies between 254–255 males and 288–289 females were used — hence the *N*s are averaged.

In addition, there is more recent evidence of "genetic cognitive decline" from a molecular genetic standpoint looking at "polygenic scores." Again this is something that will be covered in more depth in later sections. But for now - super-position this.

This is from the paper "What Caused over a Century of Decline in General Intelligence? Testing Predictions from the Genetic Selection and Neurotoxin Hypotheses"[39]. They look at genetic correlates with IQ in the general population, and then instead of tracking IQ per se, they track how different cohorts vary in those genetic correlates with IQ. Because your expressed IQ is a function of genes and environment, this just looks at the changes in genetic correlates with IQ by birth year:

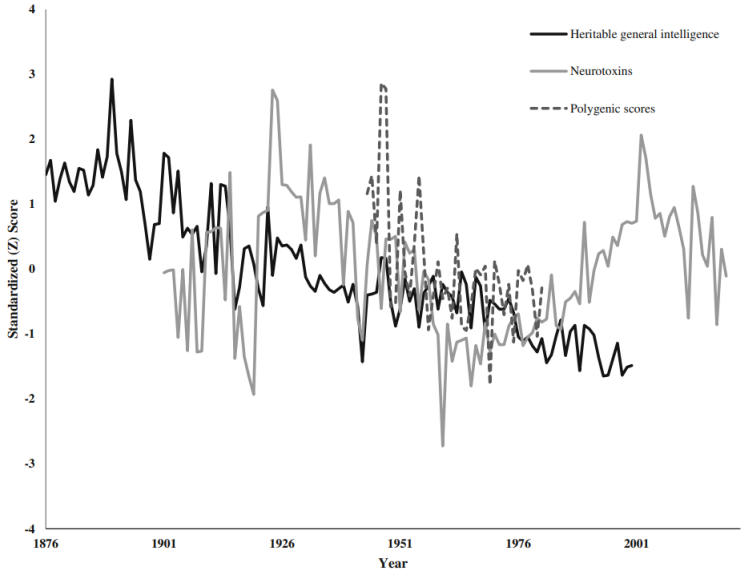


Fig. 1 Temporal trends for the neurotoxin and polygenic score chronometric factors along with *g.h* (with the predictors lagged by 25 years)

[39]https://sci-hub.tw/10.1007/s40806-017-0131-7

Now, older cohorts may have lower IQs when an IQ test is put in front of them, and until recently this was reliably true. But the genetic correlates with IQ have been declining at least since 1946 according to this, and the raw IQ scores have outright begun to decline since around the year 2000 at the latest. Which is to say the declines in genetic capacity appear to have finally caught up to effects of environmental improvements.

But from a scientific innovation standpoint, the issue is even starker. Because with your innovators, you're operating at the tails. And in a very real sense, environment isn't as important for geniuses. Environmental improvements matter more for low to mid-end of the spectrum.

From this data, the average white person (since the 1876 cohort are all white people) from 1876 would have a "genetic IQ" of around $\sim$ 130, very close to the expressed IQ of a modern academic if you go with the quantitative genetics methods (twin studies, kinship correlations, subtest heritability analyses).

If you go by the molecular genetics data (within-group validated genetic correlations), the decline appears to have gone from a modern equivalent of $\sim$ 115 in 1946, to a modern equivalent of $\sim$ 95 by the 1984 cohort. This can be a function of higher fertility among low polygenic scoring people, and / or immigration of people with lower polygenic scores.

This is an effect the eugenicists predicted, and were motivated to prevent happening.

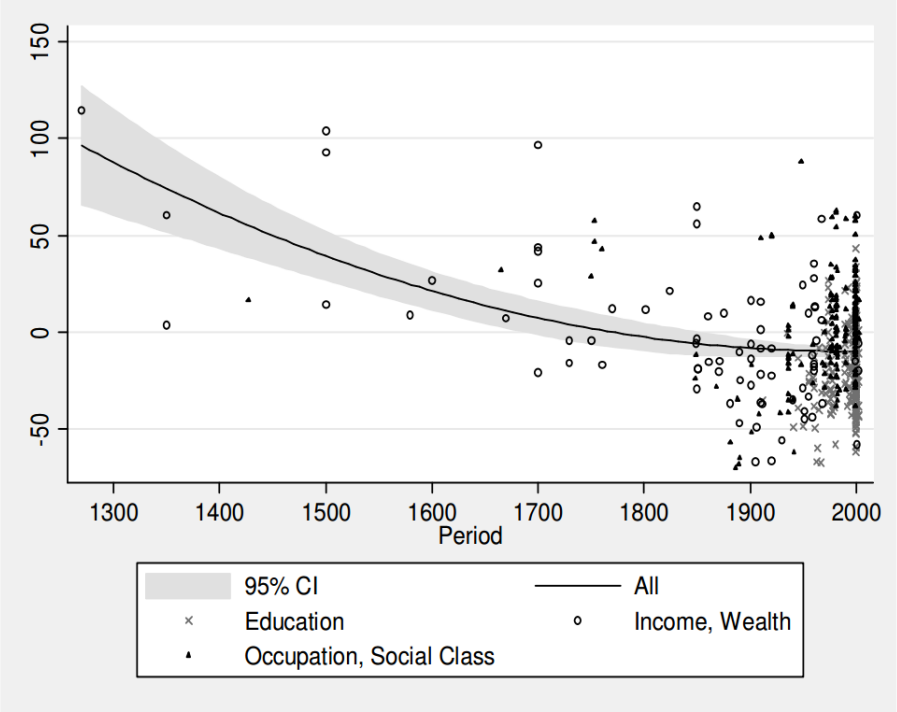
IF all of this is true, then in a sense, we are living with a level of technology that we could never create if we had to start where the people from 1876 started. And it would also explain - among other factors - why making improvements off of the current base is so difficult. The institutional changes could be both a result of lower innate intelligence (midwit managerial behavior) and a cause of scientific slowdown.

If this is overstated - then it is overstated and the effects from "genetic intelligence" decline are less stark than this chart appears. After all, polygenic scores are still a work in progress, they could still be finding non-causal correlations (which is something they try to avoid but nobody's perfect). However, it is unlikely that there has been *no* genetic decline.

And the reason for this ties into the European revolution - again, something that will be covered in depth later. Which is that there were profound changes in the breeding patterns of Europeans - most keenly felt in Britain, northern France, the Low Countries and Germany west of the Oder - that resulted in the upper-classes outbreeding the lower-classes. And from 1100-1800 AD, the death penalty was used to such an extent that approximately 1.5% of males were executed each generation - and if there was any genetic contribution to doing things that got you executed - either criminality or being stupid enough to get caught - this was a genetic culling of those traits.

However by 1850, the fertility rates of the lower classes surpassed those of the upper classes, and that is likely when the genetic decline began. From the paper "Fertility trends by social status"[40], Skirbekk et. al documents when the lower classes began outbreeding the upper classes in western Europe:

[40]https://www.demographic-research.org/volumes/vol18/5/18-5.pdf



And to the extent there's *any* relation between income, education level and "social class," another explanation of the great stagnation is the decline in cognitive ability among people in general.

Combine this with the explosion in the number of PhDs, and the value of the PhD per capita is far lower than it used to be. They are not only less elite simply because there are so many more of them, but they likely come from a dumber general population than existed in the past.

5.8 Demographic changes

This factor is more controversial and so I will merely state the coincidence.

The percentage of PhDs earned by women:

- 1976 - 23%
- 1986 - 35%
- 1996 - 40%
- 2006 - 45%
- 2017 - 53%

We can also look at college graduates that are black or hispanic as a proportion of the white college graduates:

Percent of College Graduates who are Either Black or White		
Year	White Percent	Black Percent
1940	97.14	2.86
1950	96.33	3.67
1960	94.96	5.04
1970	93.44	6.56
1980	94.06	5.94
1990	92.74	7.26
2000	90.49	9.51
2010	88.55	11.45

Percent of College Graduates who are White, Black or Hispanic			
Year	White Percent	Black Percent	Hispanic Percent
1980	91.21	5.76	3.03
1990	88.83	6.96	4.21
2000	85.22	8.96	5.82
2010	81.97	10.59	9.26

In 1994, Davidson came up with a definition of "intellectual elite", and looked at the religious affiliation of these intellectual elites. By religion, he documented the rise of Jews and Catholics and the fall of protestants among the elites from the 1930s to the 1970s, which coincides with the beginning of the great stagnation:

Religious Affiliations of Intellectual Elites: 1930-1931 and 1976-1977

Religious Groups	*1930-1931 (Fry)*	*1976-1977 (Verba and Orren)*	*Difference*	*1976-1977 as % of 1930-1931*
Liberal Protestant	52.9	27.3	-25.6	.52
Episcopalian	18.7	10.3	-8.4	.55
UCC/Congregational	14.9	4.6	-10.3	.31
Presbyterian	19.3	12.4	-6.9	.64
Moderate Protestant	20.9	19.5	-1.4	.93
Methodist	16.0	11.3	-4.7	.71
Lutheran	2.0	6.7	+4.7	3.35
Disciples of Christ	2.0	1.5	-.5	.75
Reformed	.9	—	-.9	.00
Conservative Protestant				
Baptist	8.5	5.1	-3.4	.60
Catholic	2.9	14.9	+12.0	5.14
Jewish	.9	10.3	+9.4	11.44
Other	13.9	22.6	+8.7	1.63
Unitarian-Universalist	7.0	7.2	+.2	1.03
Christian Science	.9	.5	-.4	.55
Other	6.0	14.9	+8.9	2.48
Total percent	100.0	99.7		
Total number	6,011	194		

As Jews, Catholics, women, blacks and hispanics grow in academia, this has coincided with the rise of a kind of corporate academic structure, increasing amounts of group work, review boards and an increase in the average age at which someone gains a professorship. Basically their 20s are lost and only by their mid 30s is one finally able to engage in their own research if they are lucky enough to earn a professorship. It has become more structured and less free as it becomes less male-white-anglo-saxon-protestant. These two events - the demographic change and the structural change - are perhaps completely independent of each other, perhaps related, but we know they happened coincidentally along with a decline in measured psychometric "g."

It was when academia was male-white-anglo-saxon-protestant that academia gained it's reputation for doing amazing things. This reputation persists even as those people no longer predominate it.

Charles Murray wrote a book called "Human Accomplishment" where he used scientific encyclopedias as a metric for "significant events." While such encyclopedias indeed show a "eurocentric" story of human accomplishment up to the 1990s before China became what it is today, this is not necessarily inappropriate - after all, where did the Industrial Revolution happen? Who conquered most of the world? Whatever your opinions are of the European empires, they undoubtedly reflected a great deal of competence in their achievement.

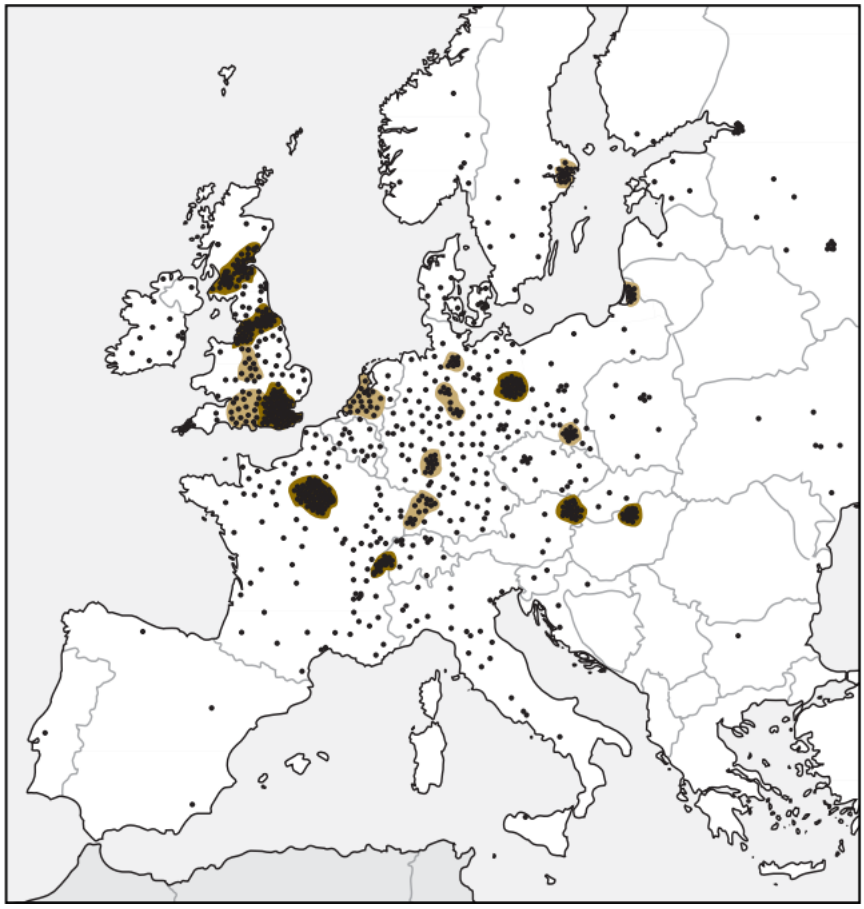
And even if you think Europeans' great wealth came from stealing from others (something which will be quantitatively falsified later), certainly the initial establishment of those empires reflected some high level of competence before any wealth had been extracted in the first place. And even if the wealth extracted made it easier to make more scientific accomplishments - well then that would vindicate a Eurocentric description of scientific advancement, even if the reasons weren't "nice."

And so a "eurocentric" view of technological progress from 1400 to 1900, and frankly up to the turn of the millenium, is not on its face an absurd point of view.

Moreover, when comparing with non-European encyclopedias, Murray found that non-Europeans were generally *more* Eurocentric than the European ones, which likely is a reflection of the fact that non-European encyclopedias are not trying to guard against the scourge of Eurocentrism, while Europeans are. Chinese and Japanese, for example, don't generally have any desire to lie about the relative technological superiority of Europe during this period, and instead focused first of emulating and perhaps surpassing it.

Perhaps in the Muslim world, there is a bias on elevating Muslim inventions. In India, on Indian inventions. But in European encyclopedias, a bias in favor of *all* non-European inventions, which would make the European encyclopedias the most biased against Europe, and only European encyclopedias would elevate African inventions since Africans generally don't make any commonly used Encyclopedias.

Within Europe itself, Murray created a map of where such significant scientific events occurred from 1800 to 1950:



Well now you have to invoke an anti-Catholic bias and an anti-Eastern Orthodox bias.

Britain (Protestant), Northern France (Protestant), Germany (Protestant), nodes in Austria and Hungary (Catholic), Holland (Protestant). This could overstate the case, with encyclopedias looking for significant events in the region where other significant events are happening, and thus missing out on significant events elsewhere. The Matthew effect.

Or perhaps this understates the disparity, as in regions and populations with lower levels of general inventiveness - any invention sticks out and is more likely to be documented. I.e. the threshold for something being considered a significant scientific event could be lower outside of this core region, resulting in an over-documentation of inventions outside of this region. Perhaps this image overstates the contributions of Spain.

Both possibilities can be speculated. And since we don't know which effect is greater, a Matthew Effect or Paul Effect, then our default assumption should be that the Encyclopedias are accurately weighing and documenting scientific advances.

Some evidence in favor of a Paul effect are the black invention myths or special sections dedicated to Islamic and Chinese inventions in Wikipedia. There is generally less focus on Spanish, Portuguese, Irish, Polish, Italian, or any Balkan inventors as these regions are part of Europe and so there is less political push to elevate these areas even though they too rank dismally in the scientific encyclopedias.

But the main point - the Universities are being filled out with people who, historically, have not

invented much. Catholics - see how sparse Spain, Portugal, Italy and Southern France are. Even Austria is less dense than the middle of Germany outside of Vienna, and Hungary has nothing outside of Budapest.

Every group, with the exception of Jews, that is increasing in the US Universities, are groups which historically have not invented much. And any credit to Jews should be tempered by their pivotal role in killing the Eugenics movement and causing you to think Eugenics to be monstrous.

Women, for example. Now, you may say that is because women lacked the opportunity. It's debatable how much of an effect gender discrimination in education had on the kind of women who would make scientific discoveries. Remember that the masses are irrelevant for this; the question is did a lack of education opportunity prevent a potential female Isaac Newton from making and proliferating a great discovery? This is not something one can just assume to be the case.

But just as a matter of fact, what's happening is an increase in the power and influence of a group (women) which hasn't invented much in the past, in hopes that they will in the future.

And the rise of these groups correlates with a changing of the basic structure of the University from the particularly independent and free-flowing structure from when a small group of protestant white males from a few countries ran the academy - to a more rigid and formal structure that resembles the Catholic church and more authoritarian societies around the rest of the world. That is to say, the University is beginning to resemble the world outside of where most things were invented from 1800-1950.

It's a gamble, and currently it looks like it's *not* paying off.

But if we know anything about professional forecasting, predictions are hard, and foxes generally beat hedgehogs. This one big thing seems to say scientific development, all else being equal, will slow down. But all else is never equal, and people who's predictions are based on one big thing - hedgehogs - are less often accurate than those whose predictions are based on lots of little things - foxes. But sometimes the hedgehog is right and the fox is wrong. More often the other way around, but it happens.

5.9 Incentive Problem

Most universities don't have any limitations on who can sit in on their classes. This is very curious: if the degree is so valuable, and it's valuable because of the knowledge held in that degree - should the university not be more guarded of its contents? Why is there not an underground market of bootleg recordings of Harvard Medical classes?

In military conflicts, intelligence is valuable precisely for its contents, and is jealously guarded. Trade secrets are jealously guarded by firms and guilds and secret societies. But the lectures of a university are not.

In fact, many elite universities are putting thousands of hours of their courses online for free. Why would they do this? Aren't they concerned that they'll stop making that tuition money once people have access to all this free knowledge?

This makes sense if the imparted knowledge isn't actually what's of value. If what matters is not the imparted knowledge, but title and rank (and some basic competence for the professions like doctors and engineers), then who cares? Engineers have to make sure the things they make work, doctors can't be totally incompetent, but otherwise - so what?

What is guarded is the formal degree title. Universities have very tight control on their records of who has what degree. They guard what is actually valuable - the title itself. The knowledge

which that title represents is less guarded and, in the sense of putting the courses online for free - often not guarded at all.

The fact that the knowledge of a student depreciates faster than the value of a car is not a problem these universities care much about, and likely don't even know about. There's a handful of studies on this matter, but if the value of the university was in the knowledge itself - this knowledge decay would be an existential crisis and scandal. But it's not. Nobody really cares. And them not caring only makes sense if the course content isn't actually very important. Because they're not selling knowledge - that's already free.

They're selling rank and title. This rank and title which explains most of the variance in how papers are evaluated and how seriously they're taken. Among people who rate a paper more highly if it includes random irrelevant math formulas and don't know what a p-value is.

5.10 Origin of the University

The origin of the University is as an outgrowth of the Catholic Church. This is something you can read up on elsewhere. Unlike the rest of this section, this point is not controversial or novel. This is not necessarily a condemnation of the University - after all, all institutions have to start from somewhere. Why not the Church? Would it be better if Universities traced their origins to pubs? (Well in our opinion that would be better and Universities would probably be less stultifying today if that were the case, but you see the point).

Originally this is because the Church needed literate clergy who could read the bible, and do basic math and accounting in order to manage the finances of a parish, and geometry to manage the Church lands. Churches would set up "Cathedral Schools", which would be courses taught by a priest on all of the things a parishioner needed to know. At first these courses would be taught either inside the Church buildings, in rented rooms, in pubs, or even outside in a field.

Over time these became more formalized, and the Church would build dedicated buildings for the education of parishioners. Eventually these schools began teaching other topics - particularly law, medicine and the natural sciences. But the same organizational structure was retained.

Even the terminology, for example the "professor", is a holdover from the Universities' origins as an adjunct to the Church.

Harvard was actually founded in America for the initial purpose of educating Clergy in the new world, since one couldn't easily travel back to England to do so.

For most of Europe's history by time - organized science was done mostly by Clergymen. The secularization of science was gradual, and the separation of the University from the Church was likewise gradual. And in many places there are Catholic high schools and colleges which are on the same grounds and part of the same legally recognized corporation as the church itself to this very day.

However, when schools were secularized - at all levels but we're most concerned with schools for those 18 and over, the "Universities" - they didn't change the formula of classes and courses and textbooks determined by a centralized body which then decided who became a Clergyman, and of what rank, or who got which kind of degree, who was allowed in - et cetera.

So while not being formally a religious body anymore - and even at ostensibly religious universities the religion aspect is usually deprecated - what exactly is the difference between the Cathedral schools and the modern University? The subject matter has changed, but Cathedral schools taught law, medicine and the natural sciences too.

The point is that the University system - supposedly the center point of science - was not scientifically constructed. It's basic structure is a tradition that dates back to Medieval Europe. And this basic structure didn't, for example, prevent the use of Lungwort to cure lung ailments because the leaf was shaped like a lung. Or prevent Galen's theory of the four humors from being the cornerstone of what passed for "psychology" for somewhere around 1,000 years.

When you get older, you begin to realize there's no such thing as an adult. Nobody really has their life together, nobody really knows what they're doing. The organizations they form - companies, schools, universities, government things - are just guys doing stuff mainly based on the tradition of that organization. There's some guy at the top, but he got there by fumbling around at the bottom for a while, or maybe he was taken from some other organization.

Academics are just people in a particular kind of organization that's been around since around 600 AD, with some minor changes. Nobody designed it, they're sitting on councils that nobody alive created or knows if they're even needed, they just kinda do it because that's what people before them did. Maybe they have a few notions that something's a little absurd about all of this, and this fundamental critique of organization is a very anarchistic sentiment and "seeing", and could be explored at it's own depth. A kind of seeing that people are all naked and these organizations are really a kind of shared hallucination among sentient apes. And that includes, especially, the university, because no such organization is held in such reverence. It has absolutely replaced the Church's role as the last stop on questions of truth.

And all this explains why you have far more intellectual freedom in a bowling alley than a university.

5.11 Organizations

When thinking about evolution, most tend to only think about it at the level of the organism. But organizations themselves can be seen as a kind of organism. Now we're not saying that Chick-Fil-A has qualia or it's own mental states or anything like that, but organizations can change, they can grow, shrink, die. The organizations that exist today, what do they all have in common? They exist.

The organizations that exist today are those which are good at existing. It doesn't have to be some conspiracy among people (though "conspiracy" shouldn't be a dirty word, they happen all the time).

Cells compete with each other, and one viable way for cells to compete was to group up and form tissues. Then multiple tissues glob together to form organs, then systems of organs or "organ systems." At this point, organisms are competing with each other and evolution occurs between the organisms, often to the detriment of their component cells.

For example, an organism may have 1 million members. It then evolves to be smaller - $\frac{1}{4}$ the size, but the species doubles in population. Twice as many organisms! But half as many cells... so the organism evolves based on the requirements of the organism, not on it's component cells. The cells become subjugated to the evolutionary pressures of the organism.

The same can occur for an organization. A cult that forces people to make wasteful sacrifices taps into the "sunk cost fallacy". Where people are unwilling to stop doing and/or believing something after they have made some sacrifice for it. They make sacrifices, which reduces the reproductive abilities of the organisms in the cult, but makes them more committed to the cult because of the sunk cost fallacy.

"I spent 10 years getting my masters degree and was thus unable to have children. However, given that sacrifice, I must now rationalize it and support the institution I was committed to, because to admit this as folly would be psychologically catastrophic."

"I spent $200,000 worth of time and money getting an undergraduate degree. I will now value that degree personally and in my hiring decisions because admitting it was worthless would be too much of a revelation to bear."

And so the cycle can continue, the organization can persist even if it reduces the reproductive and economic vitality of the host (you, and the rest of society). We're not saying these thought experiments, mere words, prove that this is in fact the case - though we *think* it is. We're merely saying that it could be, and the persistence of an organization is only evidence that it is good at persisting, *not* that it does any good at the organism level.

An organism will only evolve to increase the number cells *if* that also helps to increase the number of organisms. An organization will only evolve to increase either the reproductive success or economic success of the organisms in the organization IF promoting such success also helps the organization. That's all. The organizations around you have stated functions - after all, organisms must promote their existence. If organisms, humans, don't continue to give life to these organizations, they can dissolve these organizations. And so organizations usually - usually - must convince humans that they are necessary and/or beneficial. They don't have to actually be necessary or beneficial, they just must usually convince the right people that they are.

Or do they? Because say firms hire people with college degrees, even if those degrees don't predict better employee performance. The organization can continue this policy even if only 10% of the people in the firm agree with that policy of preferring to hire people with college degrees. So long as those 10% are the policy-makers and everyone else just goes along with it. Heck it could be 1%. Or it could be none if, for whatever reason, firms don't change this policy even if zero people believe in it.

That would probably be harder, but a policy could just be in place for a long time, nobody agrees with it, but nobody is willing to organize within the organization enough to change it. And with all of the people acting individually, their individual best-action is, so they think, to get a degree.

This would be enough to sustain people going to colleges and getting inherently worthless degrees, but valuable to them individually because firms pay them more for it. Even if nobody actually supports this arrangement. So the organization, and an organizational policy can theoretically persist even if nobody supports it and it does no good for humans as a whole.

In reality though some people do support the university degree system and more importantly the traditions of firms taking it into consideration.

The punchline: don't cite institutions supporting something as evidence for it. There are all manner of reasons an institution can exist and support a policy or position that has nothing to do with truth or what's good for humans, or any particular group of humans.

5.12 Scam

Let me tell you about this crazy cult. This cult takes people when they're 18, isolates them from the rest of society for years - 4 years for the first level, 6 for the second, 8 for the third *at minimum*. In practice, people typically have to spend longer to reach these levels.

They also have to give this cult a bunch of money for the "privilege" of doing so. This cult has wormed it's way into government and corporations - creating a job and client network. At all

manner of positions, you can only get certain jobs if you are at a certain level in the order. None of this is based on employee performance data - people who have a "degree" in geology can work at an IT firm. What matters most is the level itself.

And strangely enough, it doesn't vary by subject the member chooses to get a "degree" in. An English degree takes precisely as long to get as a Math degree. What a coincidence that is!

Most cults demand their adherents do strange things, believe strange things, and pay them money. Now, what is a "strange thing" or "false belief" is something that takes a very long time to sort out. You can say "I don't believe in Xenu or the soul catchers", but actually disproving that is a lot of work.

But what we can do is look into the time and financial commitments of cults. One of the most infamous cults is Scientology. Now based on interviews and anecdotes, you may get the impression that going through the entire bridge to "Clear" and the "OT levels" costs millions. But this comes from former scientologists who have chosen to publicly speak out against it. And even if you think scientology is awful, certainly our impression of it is very negative, one must take such stories with a grain of salt.

To reach the state of "dianetic clear," one must pay roughly ∼$140,000 to go clear. That is the base "tuition" cost. This is a lot, but how much does an undergraduate degree cost? Well it depends on the university. For elite private universities, $160,000 for tuition is not unheard of.

Scientology is also known for having all manner of hidden costs and "sec checks" and "auditing." But Universities likewise have all manner of required fees on top of the tuition. Which is greater? Well, ex-scientologists who speak out against the church would have you believe these fees end up costing more than the course itself. Perhaps for them it did, perhaps they're lying or misremembering, who knows.

What if your only knowledge of university fees was what critics and former "Universitarians" told you? Well they'd certainly tell you the extra costs were higher than they actually were.

Universities have scholarships, but in fairness, Scientology also has variable pricing - people with more money pay more.

But also, keep in mind that universities get money from government and external organizations for their "scholarships." Scientology doesn't have access to this, so they can appear less generous simply because they don't have access to the broad web of societal support that Universities do.

You can respond that people with college degrees end up with higher lifetime earnings. And that's true. But what if scientology had the same social status as the Universities? Then they would be able to say that "going clear" leads to higher lifetime earnings, as businesses would only hire people who are "clear", and certain positions would require various ranks of Operating Thetan (equivalent to graduate and post-graduate degrees in Universitarianism).

That said, as a scientologist, you have access to the Scientologist business network. It's not known how much this is worth. Certainly it's much smaller, but it also has far fewer people competing for it. But if Scientology had the level of legitimacy that Universities have, then they could absolutely say that "going clear" will lead to you earning more money.

So higher earnings cannot be an argument for universities being inherently less of a cult than Scientology. These are just the effects of broader societal acceptance - things Scientology would have if it had broad societal acceptance.

What about all of the stories about abuses in Scientology? Well, again, stories. If your only knowledge of Universities came from critics of them, people who claim to have been abused, you'd likely think of Universities the way you think of Scientology.

Universities teach useful things though! Well, virtually all ex-scientologists claim that some of the stuff they learned in scientology were useful and helped them get over their emotional problems. Helped them be less fearful of conflict, helped them be more confident in asking people out. Similarly, people with degrees in geology working for some IT company may point out that they too learned some things in college that help them today.

What about research scientists? Well, if Scientology took over the role of the University, no doubt they would teach all the research scientists, geneticists, engineers, etc. In fact, even today in it's rump state compared to the University, Scientology has trade schools. Similar to how the Catholic Church went from just teaching an ideological doctrine to teaching Law and Medicine and the Natural Sciences. But the Catholic Church didn't have to compete with the modern University system or have the internet exposing the ridiculous things in its texts (which the Bible has no shortage of).

What about the time commitment? Well Scientology seems to be more flexible, as people can take the courses as their schedule permits, versus the traditional 4 years (optimistically) on campus. This is changing though. However, keep in mind that the reputation of Universities was established before more flexible hours existed.

In terms of time commitment - who knows? The time commitment for becoming an "Operating Thetan" is much higher than going clear, but the time commitment for a PhD is also higher than getting an undergraduate degree.

What about abuses? Particularly among the "Sea Org" you hear all manner of horror stories. But among graduate and post-graduate students, they are similarly forced to work long hours with extremely low pay, made to work in low-paying or even unpaid internships. Again, if the only thing you knew about graduate and post-grad work was what critics and ex-"Universitarians" had to say, you might hear horror stories similar to those of ex-Scientologists.

And if asked to respond, Universitarians would likely sound like Scientology PR in denying these abuses.

Scientologists are generally more angry in their response to non-scientologist critics, but what if Universities had the same general societal scorn that Scientology does? Perhaps they would likewise become somewhat unhinged under the pressure. Anyone who has been mobbed can certainly understand this.

The point here is not to say "Scientology is good," we certainly don't believe that. But give the devil his due; many of the criticisms of scientology are for things that the university does and is generally accepted. And they are things which if the University did not have widespread perceived legitimacy, would be considered cultish and abusive.

In addition, if you were to go back in time to when the University was much smaller, which is to say before the term "university" even existed and it was just the Church with informal Cathedral schools, and these schools were set up in China and had to compete with already-established pre-existing Chinese Universities, while promoting a doctrine completely alien to Chinese beliefs - well, how do you think it would be perceived? Imagine how many lies would be told about it. And how is Scientology doing?

Now we think Scientology is literally insane. But so is Christianity, and so is the converged yet amorphous doctrine of "The Modern University."

A key difference is that the insanity of Scientology and Christianity was bound by sacred texts. But what exists today is the same theological structure of the Cathedral schools in "secular" universities, but with a more free-flowing set of beliefs. The same tyrannical structure of the

old Church is in place, one which existed to institutionalize and enforce doctrine, and then got jury-rigged to teach other subjects.

But now the "correct" beliefs are constantly morphing. In one sense the University today is better than the old Church as people have far less tolerance for hardship, and therefore violence, than they did in the past. But in another sense it's worse because it's anarcho-tyranny; things that were acceptable 4 years ago are now things that get you excommunicated - not by a pope or council, but by a mob. And there's no clear criteria for heresy or blasphemy.

Such radical changes in doctrine were extremely rare before 1900, and matters of heresy and blasphemy were clearly delineated. Incidentally, this is why people with deviant views in the US today often consider China's "Social Credit" system less tyrannical than "cancel culture." Because the Social Credit system, being a government operation, has a kind of explicitness and rationality of process. It is clear what is and isn't a violation - and violations can be legally remedied, which firms are then legally bound to recognize.

With "cancel culture" there is no formal path to remedy, what constitutes a cancellable offense is ambiguous and varies based on who you are, and firms are not legally bound to respect any remedy.

Of course we absolutely don't support the Chinese Social Credit system, but it is undeniably less tyrannical than the situation in the United States.

Now this might be more difficult for you since you exist in the context of your place and time in history, but the University teaches things that are, in our opinion, also literally insane. Things like the denial of gender, the idea "white supremacy" exists, the idea that human populations don't differ in genetic predispositions to intelligence. And these are things that get you excommunicated for denying. As they should be right? Racists, Nazis, Fascists! Get that blood boiling, burn the heretic!

Right now this probably goes too far for you. It's one thing to show the Church is corrupt; that's the easiest thing to do. In fact recognizing the corruption of the church is so common that a kind of second schism occurred in Europe against the Catholic Church. They were still Christians, but broke off from the Pope.

In Scientology, this same process is happening writ small. They're called "Independent Scientology," or "squirrels" by the Church of Scientology. They practice Scientology while not being part of the Church of Scientology. They still believe in Scientology and the teachings of ~~The Authors of the Gospels~~ ~~Muhammad~~ ~~Joseph Smith~~ L. Ron Hubbard but oppose the perceived abuses of the Church. And while a scientific analysis of this hasn't been done to our knowledge, in interviews of former scientologists, the *first* thing they mention as part of their road to leaving are the abuses of the Church. In fact Tori Magoo, a popular ex-scientologist, mentions that she still believed in Scientology when she left the church.

Of course since then she went on about the dangerous cult of George Bush when that hysteria was popular. Which also ties into research that Scientologists are inherently "liberal." Meaning that if they weren't in the Church of scientology they would otherwise be "liberals" and that people who join cults either lean "left" or would otherwise lean "left" if they weren't in a cult. (These are terms we don't endorse, but we're using them for now in the absence of an alternative we will present later.)

And we have no illusions that you will stop being a "progressive" or "liberal" or whatever self-congratulatory label you use for beliefs converged with the University - just because you have in a sense "left" or at least lost faith in the institution which was necessary to convince you and society

of that ideological content. But it's a necessary first step, because if you have lost that faith, well, now magic science man is gone. You can't just appeal to the wizards in their tower to know all the reasons that evolution stopped at the neck. Either you have to know them, or accept that divergent evolution applies to humans and it applies to the brain, and that your "racist" instincts aren't evil but actually evolved to protect you from annihilation.